2 *Probability*

2.1 Introduction

> Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

In the previous chapter, we saw how probability can play a useful role in machine learning. In this chapter, we discuss probability theory in more detail. We do not have space to go into great detail — for that, you are better off consulting some of the excellent textbooks available on this topic, such as (Jaynes 2003; Bertsekas and Tsitsiklis 2008; Wasserman 2004). But we will briefly review many of the key ideas you will need in later chapters.

Before we start with the more technical material, let us pause and ask: what is probability? We are all familiar with phrases such as "the probability that a coin will land heads is 0.5". But what does this mean? There are actually at least two different interpretations of probability. One is called the **frequentist** interpretation. In this view, probabilities represent long run frequencies of events. For example, the above statement means that, if we flip the coin many times, we expect it to land heads about half the time.[1]

The other interpretation is called the **Bayesian** interpretation of probability. In this view, probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to information rather than repeated trials (Jaynes 2003; Lindley 2006). In the Bayesian view, the above statement means we believe the coin is equally likely to land heads or tails on the next toss.

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might want to compute the probability that the polar ice cap will melt by 2020 CE. This event will happen zero or one times, but cannot happen repeatedly. Nevertheless, we ought to be able to quantify our uncertainty about this event; based on how probable we think this event is, we will (hopefully!) take appropriate actions (see Section 5.7 for a discussion of optimal decision making under uncertainty). To give some more machine learning oriented examples, we might have received a specific email message, and want to compute the probability it is spam. Or we might have observed a "blip" on our radar screen, and want to compute the probability distribution over the location of the corresponding target (be it a bird, plane, or missile). In all these cases, the

1. Actually, the Stanford statistician (and former professional magician) Persi Diaconis has shown that a coin is about 51% likely to land facing the same way up as it started, due to the physics of the problem (Diaconis et al. 2007).

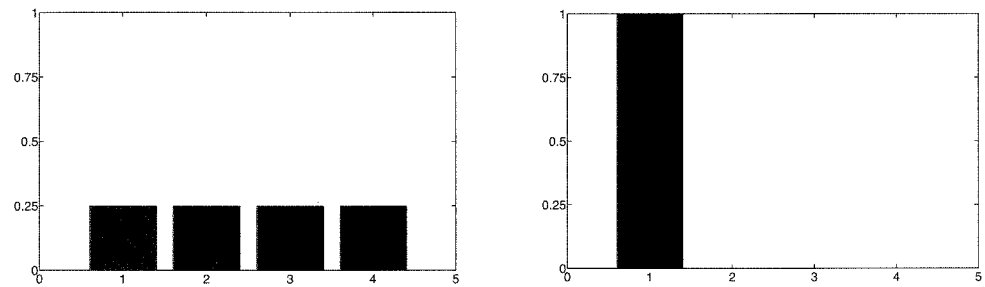


Figure 2.1 (A) a uniform distribution on $\{1, 2, 3, 4\}$, with $p(x = k) = 1/4$. (b) a degenerate distribution $p(x) = 1$ if $x = 1$ and $p(x) = 0$ if $x \in \{2, 3, 4\}$. Figure generated by `discreteProbDistFig`.

idea of repeated trials does not make sense, but the Bayesian interpretation is valid and indeed quite natural. We shall therefore adopt the Bayesian interpretation in this book. Fortunately, the basic rules of probability theory are the same, no matter which interpretation is adopted.

2.2 A brief review of probability theory

This section is a very brief review of the basics of probability theory, and is merely meant as a refresher for readers who may be "rusty". Please consult some of the excellent textbooks available on this topic, such as (Rice 1995; Jaynes 2003; Bertsekas and Tsitsiklis 2008; Wasserman 2004; Lindley 2006), for further details, if necessary. Readers who are already familiar with these basics may safely skip this section.

2.2.1 Discrete random variables

The expression $p(A)$ denotes the probability that the event A is true. For example, A might be the logical expression "it will rain tomorrow". We require that $0 \leq p(A) \leq 1$, where $p(A) = 0$ means the event definitely will not happen, and $p(A) = 1$ means the event definitely will happen. We write $p(\overline{A})$ to denote the probability of the event not A; this is defined to $p(\overline{A}) = 1 - p(A)$. We will often write $A = 1$ to mean the event A is true, and $A = 0$ to mean the event A is false.

We can extend the notion of binary events by defining a **discrete random variable** (**rv** for short) X, which can take on any value from a finite or countably infinite set $\mathcal{X}$. We denote the probability of the event that $X = x$ by $p(X = x)$, or just $p(x)$ for short. Here $p()$ is called a **probability mass function** or **pmf**. This satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. Figure 2.1 shows two pmf's defined on the finite **state space** $\mathcal{X} = \{1, 2, 3, 4\}$. On the left we have a uniform distribution, $p(x) = 1/4$, and on the right, we have a degenerate distribution, $p(x) = \mathbb{I}(x = 1)$, where $\mathbb{I}()$ is the binary **indicator function**. This distribution represents the fact that X is always equal to the value 1, in other words, it is a constant.

2.2.2 Fundamental rules

In this section, we review the basic rules of probability.

2.2.2.1 Probability of a union of two events

Given two events, A and B, we define the probability of A or B as follows:

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B) \tag{2.1}$$

$$= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \tag{2.2}$$

2.2.2.2 Joint probabilities

We define the probability of the joint event A and B as follows:

$$p(A, B) = p(A \wedge B) = p(A|B)p(B) \tag{2.3}$$

This is sometimes called the **product rule**. Given a **joint distribution** on two events $p(A, B)$, we define the **marginal distribution** as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b) \tag{2.4}$$

where we are summing over all possible states of B. We can define $p(B)$ similarly. This is sometimes called the **sum rule** or the **rule of total probability**.

The product rule can be applied multiple times to yield the **chain rule** of probability:

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3) \ldots p(X_D|X_{1:D-1}) \tag{2.5}$$

where we introduce the Matlab-like notation $1 : D$ to denote the set $\{1, 2, \ldots, D\}$.

2.2.2.3 Conditional probability

We define the **conditional probability** of event A, given that event B is true, as follows:

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0 \tag{2.6}$$

2.2.3 Bayes' rule

Combining the definition of conditional probability with the product and sum rules yields **Bayes' rule**, also called **Bayes' Theorem**[2]:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')} \tag{2.7}$$

Sir Harold Jeffreys wrote that "Bayes's theorem is to the theory of probability what Pythagoras's theorem is to geometry" (Jeffreys 1973). We show two simple applications of the theorem below, but we will encounter many, many more throughout the book. (See also (McGrayne 2011) for an interesting historical account of its applications to many other real-world problems.)

2. Thomas Bayes (1702–1761) was an English mathematician and Presbyterian minister. Technically speaking, we should write "Bayes' rule", but often we will just write "Bayes rule" instead, dropping the apostrophe, since it is less fussy.

2.2.3.1 Example: medical diagnosis

As an example of how to use this rule, consider the following medical diagnosis problem. Suppose you are a woman in your 40s, and you decide to have a medical test for breast cancer called a **mammogram**. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is. Suppose you are told the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8. In other words,

$$p(x = 1|y = 1) = 0.8 \tag{2.8}$$

where $x = 1$ is the event the mammogram is positive, and $y = 1$ is the event you have breast cancer. Many people conclude they are therefore 80% likely to have cancer. But this is false! It ignores the prior probability of having breast cancer, which fortunately is quite low:

$$p(y = 1) = 0.004 \tag{2.9}$$

Ignoring this prior is called the **base rate fallacy**. We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$p(x = 1|y = 0) = 0.1 \tag{2.10}$$

Combining these three terms using Bayes rule, we can compute the correct answer as follows:

$$p(y = 1|x = 1) = \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \tag{2.11}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 \tag{2.12}$$

where $p(y = 0) = 1 - p(y = 1) = 0.996$. In other words, if you test positive, you only have about a 3% chance of actually having breast cancer![3]

2.2.3.2 Example: Generative classifiers

We can generalize the medical diagnosis example to classify feature vectors $\mathbf{x}$ of arbitrary type as follows:

$$p(y = c|\mathbf{x}) = \frac{p(y = c)p(\mathbf{x}|y = c)}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c')} \tag{2.13}$$

This is called a **generative classifier**, since it specifies how to generate the data using the **class-conditional density** $p(\mathbf{x}|y = c)$ and the class prior $p(y = c)$. We discuss such models in detail in Chapters 3 and 4. An alternative approach is to directly fit the class posterior, $p(y = c|\mathbf{x})$; this is known as a discriminative classifier. We discuss the pros and cons of the two approaches in Section 8.6.

3. These numbers are from (McGrayne 2011, p257). Based on this analysis, the US government decided not to recommend annual mammogram screening to women in their 40s: the number of false alarms would cause needless worry and stress amongst women, and result in unnecesssary, expensive, and potentially harmful followup tests. See Section 5.7 for the optimal way to trade off risk versus reward in the face of uncertainty.

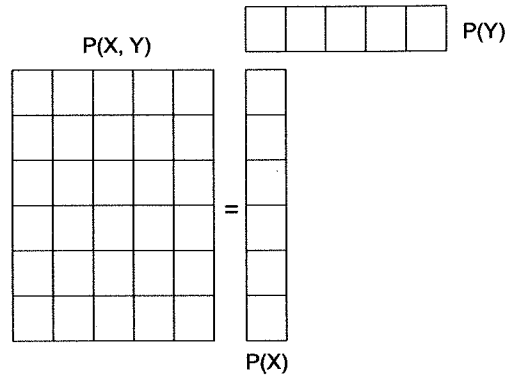


Figure 2.2 Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$. Here X and Y are discrete random variables; X has 6 possible states (values) and Y has 5 possible states. A general joint distribution on two such variables would require $(6 \times 5) - 1 = 29$ parameters to define it (we subtract 1 because of the sum-to-one constraint). By assuming (unconditional) independence, we only need $(6 - 1) + (5 - 1) = 9$ parameters to define $p(x, y)$.

2.2.4 Independence and conditional independence

We say X and Y are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals (see Figure 2.2), i.e.,

$$X \perp Y \iff p(X, Y) = p(X)p(Y) \tag{2.14}$$

In general, we say a set of variables is mutually independent if the joint can be written as a product of marginals.

Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct. We therefore say X and Y are **conditionally independent** (CI) given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{2.15}$$

When we discuss graphical models in Chapter 10, we will see that we can write this assumption as a graph $X - Z - Y$, which captures the intuition that all the dependencies between X and Y are mediated via Z. For example, the probability it will rain tomorrow (event X) is independent of whether the ground is wet today (event Y), given knowledge of whether it is raining today (event Z). Intuitively, this is because Z "causes" both X and Y, so if we know Z, we do not need to know about Y in order to predict X or vice versa. We shall expand on this concept in Chapter 10.

Another characterization of CI is this:

Theorem 2.2.1. *$X \perp Y|Z$ iff there exist functions g and h such that*

$$p(x, y|z) = g(x, z)h(y, z) \tag{2.16}$$

for all x, y, z such that $p(z) > 0$.

See Exercise 2.8 for the proof.

CI assumptions allow us to build large probabilistic models from small pieces. We will see many examples of this throughout the book. In particular, in Section 3.5, we discuss naive Bayes classifiers, in Section 17.2, we discuss Markov models, and in Chapter 10 we discuss graphical models; all of these models heavily exploit CI properties.

2.2.5 Continuous random variables

So far, we have only considered reasoning about uncertain discrete quantities. We will now show (following (Jaynes 2003, p107)) how to extend probability to reason about uncertain continuous quantities.

Suppose X is some uncertain continuous quantity. The probability that X lies in any interval $a \leq X \leq b$ can be computed as follows. Define the events $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$. We have that $B = A \vee W$, and since A and W are mutually exclusive, the sum rules gives

$$p(B) = p(A) + p(W) \tag{2.17}$$

and hence

$$p(W) = p(B) - p(A) \tag{2.18}$$

Define the function $F(q) \triangleq p(X \leq q)$. This is called the **cumulative distribution function** or **cdf** of X. This is a monotonically non-decreasing function. See Figure 2.3(a) for an example. Using this notation we have

$$p(a < X \leq b) = F(b) - F(a) \tag{2.19}$$

Now define $f(x) = \frac{d}{dx}F(x)$ (we assume this derivative exists); this is called the **probability density function** or **pdf**. See Figure 2.3(b) for an example. Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

$$p(a < X \leq b) = \int_a^b f(x)dx \tag{2.20}$$

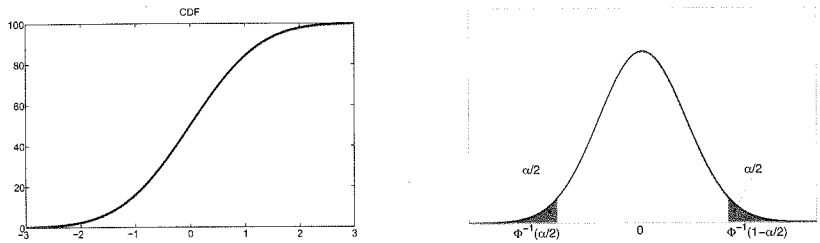
As the size of the interval gets smaller, we can write

$$p(x \leq X \leq x + dx) \approx p(x)dx \tag{2.21}$$

We require $p(x) \geq 0$, but it is possible for $p(x) > 1$ for any given x, so long as the density integrates to 1. As an example, consider the **uniform distribution** $\text{Unif}(a, b)$:

$$\text{Unif}(x|a, b) = \frac{1}{b-a}\mathbb{I}(a \leq x \leq b) \tag{2.22}$$

If we set $a = 0$ and $b = \frac{1}{2}$, we have $p(x) = 2$ for any $x \in [0, \frac{1}{2}]$.



(a) (b)

Figure 2.3 (a) Plot of the cdf for the standard normal, $\mathcal{N}(0,1)$. (b) Corresponding pdf. The shaded regions each contain $\alpha/2$ of the probability mass. Therefore the nonshaded region contains $1-\alpha$ of the probability mass. If the distribution is Gaussian $\mathcal{N}(0,1)$, then the leftmost cutoff point is $\Phi^{-1}(\alpha/2)$, where Φ is the cdf of the Gaussian. By symmetry, the rightmost cutoff point is $\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2)$. If $\alpha = 0.05$, the central interval is 95%, and the left cutoff is -1.96 and the right is 1.96. Figure generated by `quantileDemo`.

2.2.6 Quantiles

If the cdf F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}. If F is the cdf of X, then $F^{-1}(\alpha)$ is the value of x_α such that $p(X \leq x_\alpha) = \alpha$; this is called the α **quantile** of F. The value $F^{-1}(0.5)$ is the **median** of the distribution, with half of the probability mass on the left, and half on the right. The values $F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the lower and upper **quartiles**.

We can also use the inverse cdf to compute **tail area probabilities**. For example, if Φ is the cdf of the Gaussian distribution $\mathcal{N}(0,1)$, then points to the left of $\Phi^{-1}(\alpha/2)$ contain $\alpha/2$ of the probability mass, as illustrated in Figure 2.3(b). By symmetry, points to the right of $\Phi^{-1}(1-\alpha/2)$ also contain $\alpha/2$ of the mass. Hence the central interval $(\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2))$ contains $1-\alpha$ of the mass. If we set $\alpha = 0.05$, the central 95% interval is covered by the range

$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96) \tag{2.23}$$

If the distribution is $\mathcal{N}(\mu, \sigma^2)$, then the 95% interval becomes $(\mu - 1.96\sigma, \mu + 1.96\sigma)$. This is sometimes approximated by writing $\mu \pm 2\sigma$.

2.2.7 Mean and variance

The most familiar property of a distribution is its **mean**, or **expected value**, denoted by μ. For discrete rv's, it is defined as $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\, p(x)$, and for continuous rv's, it is defined as $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\, p(x) dx$. (If $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} |x|\, p(x) dx$ is not finite, the mean is not defined; we will see some examples of this later.)

The **variance** is a measure of the "spread" of a distribution, denoted by σ^2. This is defined

as follows:

$$\text{var}\left[X\right] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx \tag{2.24}$$

$$= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}\left[X^2\right] - \mu^2 \tag{2.25}$$

from which we derive the useful result

$$\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2 \tag{2.26}$$

The **standard deviation** is defined as

$$\text{std}\left[X\right] \triangleq \sqrt{\text{var}\left[X\right]} \tag{2.27}$$

This is useful since it has the same units as X itself.

2.3 Some common discrete distributions

In this section, we review some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

2.3.1 The binomial and Bernoulli distributions

Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim \text{Bin}(n, \theta)$. The pmf is given by

$$\text{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{2.28}$$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!} \tag{2.29}$$

is the number of ways to choose k items from n (this is known as the **binomial coefficient**, and is pronounced "n choose k"). See Figure 2.4 for some examples of the binomial distribution. This distribution has the following mean and variance:

$$\text{mean} = n\theta, \quad \text{var} = n\theta(1-\theta) \tag{2.30}$$

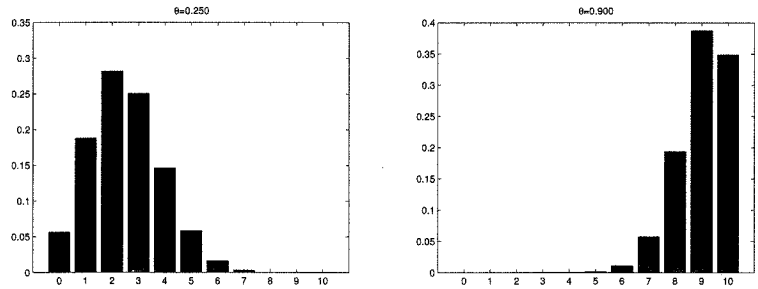
Now suppose we toss a coin only once. Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" or "heads" of θ. We say that X has a **Bernoulli** distribution. This is written as $X \sim \text{Ber}(\theta)$, where the pmf is defined as

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \tag{2.31}$$

In other words,

$$\text{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases} \tag{2.32}$$

This is obviously just a special case of a Binomial distribution with $n = 1$.



(a) (b)

Figure 2.4 Illustration of the binomial distribution with $n = 10$ and $\theta \in \{0.25, 0.9\}$. Figure generated by `binomDistPlot`.

2.3.2 The multinomial and multinoulli distributions

The binomial distribution can be used to model the outcomes of coin tosses. To model the outcomes of tossing a K-sided die, we can use the **multinomial** distribution. This is defined as follows: let $\mathbf{x} = (x_1, \ldots, x_K)$ be a random vector, where x_j is the number of times side j of the die occurs. Then $\mathbf{x}$ has the following pmf:

$$\text{Mu}(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \tag{2.33}$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!} \tag{2.34}$$

is the **multinomial coefficient** (the number of ways to divide a set of size $n = \sum_{k=1}^{K} x_k$ into subsets with sizes x_1 up to x_K).

Now suppose $n = 1$. This is like rolling a K-sided dice once, so $\mathbf{x}$ will be a vector of 0s and 1s (a bit vector), in which only one bit can be turned on. Specifically, if the dice shows up as face k, then the k'th bit will be on. In this case, we can think of x as being a scalar categorical random variable with K states (values), and $\mathbf{x}$ is its **dummy encoding**, that is, $\mathbf{x} = [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K)]$. For example, if $K = 3$, we encode the states 1, 2 and 3 as $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. This is also called a **one-hot encoding**, since we imagine that only one of the K "wires" is "hot" or on. In this case, the pmf becomes

$$\text{Mu}(\mathbf{x}|1, \boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)} \tag{2.35}$$

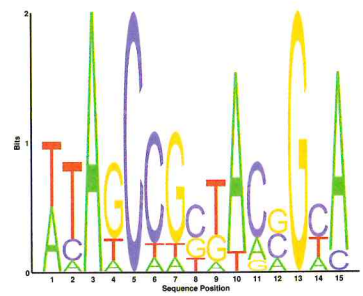
See Figure 2.1 for some examples. This very common special case is known as a **categorical** or **discrete** distribution. (Gustavo Lacerda suggested we call it the **multinoulli distribution**, by analogy with the Binomial/ Bernoulli distinction, a term which we shall adopt in this book.) We

Name	n	K	x
Multinomial	-	-	$\mathbf{x} \in \{0, 1, \ldots, n\}^K$, $\sum_{k=1}^K x_k = n$
Multinoulli	1	-	$\mathbf{x} \in \{0, 1\}^K$, $\sum_{k=1}^K x_k = 1$ (1-of-K encoding)
Binomial	-	1	$x \in \{0, 1, \ldots, n\}$
Bernoulli	1	1	$x \in \{0, 1\}$

Table 2.1 Summary of the multinomial and related distributions.

```
a t a g c c g g t a c g g c a
t t a g c t g c a a c c g c a
t c a g c c a c t a g a g c a
a t a a c c g c g a c c g c a
t t a g c c g c t a a g g t a
t a a g c c t c g t a c g t a
t t a g c c g t t a c g g c c
a t a t c c g g t a c a g t a
a t a g c a g g t a c c g a a
a c a t c c g t g a c g g a a
```

(a)



(b)

Figure 2.5 (a) Some aligned DNA sequences. (b) The corresponding sequence logo. The vertical axis represents $2 - H$, where H is the entropy of the distribution for that column (measured in bits). Thus deterministic distributions (with an entropy of 0) have height 2, and uniform distributions (with an entropy of 2) have height 0. Figure generated by `seqlogoDemo`.

will use the following notation for this case:

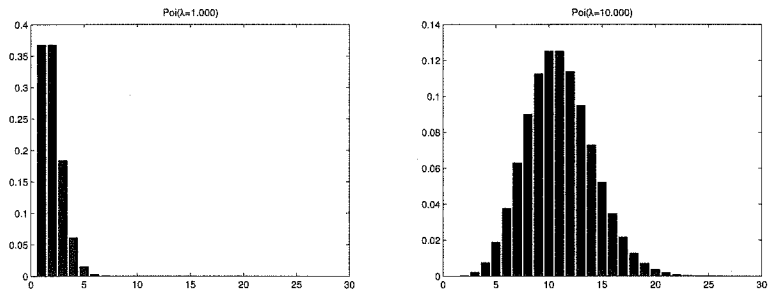
$$\text{Cat}(x|\boldsymbol{\theta}) \triangleq \text{Mu}(\mathbf{x}|1, \boldsymbol{\theta}) \tag{2.36}$$

In otherwords, if $x \sim \text{Cat}(\boldsymbol{\theta})$, then $p(x = j|\boldsymbol{\theta}) = \theta_j$. See Table 2.1 for a summary.

2.3.2.1 Application: DNA sequence motifs

An interesting application of multinomial models arises in **biosequence analysis**. Suppose we have a set of (aligned) DNA sequences, such as in Figure 2.5(a), where there are 10 rows (sequences) and 15 columns (locations along the genome). We see that several locations are conserved by evolution (e.g., because they are part of a gene coding region), since the corresponding columns tend to be "pure". For example, column 13 is all G's.

One way to visually summarize the data is by using a **sequence logo**: see Figure 2.5(b). We plot the letters A, C, G and T, with the most probable letter on the top, and with a fontsize related to their empirical probability (for details of the vertical scaling, see Section 2.8.1). The empirical probability distribution at location t, $\hat{\boldsymbol{\theta}}_t$, is obtained by normalizing the vector of



(a) (b)

Figure 2.6 Illustration of some Poisson distributions for $\lambda \in \{1, 10\}$. We have truncated the x-axis to 30 for clarity, but the support of the distribution is over all the non-negative integers. Figure generated by `poissonPlotDemo`.

counts (see Equation 3.48):

$$\mathbf{N}_t = \left(\sum_{i=1}^{N} \mathbb{I}(X_{it} = 1), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 2), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 3), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 4) \right) \tag{2.37}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{N}_t / N \tag{2.38}$$

This distribution is known as a **motif**. We can also compute the most probable letter in each location; this is called the **consensus sequence**.

2.3.3 The Poisson distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim \text{Poi}(\lambda)$, if its pmf is

$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \tag{2.39}$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents. See Figure 2.6 for some plots.

2.3.4 The empirical distribution

Given a set of data, $\mathcal{D} = \{x_1, \ldots, x_N\}$, we define the **empirical distribution**, also called the **empirical measure**, as follows:

$$p_{\text{emp}}(A) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A) \tag{2.40}$$

where $\delta_x(A)$ is the **Dirac measure**, defined by

$$\delta_x(A) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases} \tag{2.41}$$

In general, we can associate "weights" with each sample:

$$p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x) \tag{2.42}$$

where we require $0 \leq w_i \leq 1$ and $\sum_{i=1}^{N} w_i = 1$. We can think of this as a histogram, with "spikes" at the data points x_i, where w_i determines the height of spike i. This distribution assigns 0 probability to any point not in the data set.

2.4 Some common continuous distributions

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

2.4.1 Gaussian (normal) distribution

The most widely used distribution in statistics and machine learning is the Gaussian or normal distribution. Its pdf is given by

$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{2.43}$$

Here $\mu = \mathbb{E}[X]$ is the mean (and mode), and $\sigma^2 = \text{var}[X]$ is the variance. $\sqrt{2\pi\sigma^2}$ is the normalization constant needed to ensure the density integrates to 1 (see Exercise 2.11).

We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that $p(X = x) = \mathcal{N}(x|\mu, \sigma^2)$. If $X \sim \mathcal{N}(0, 1)$, we say X follows a **standard normal** distribution. See Figure 2.3(b) for a plot of this pdf; this is sometimes called the **bell curve**.

We will often talk about the **precision** of a Gaussian, by which we mean the inverse variance: $\lambda = 1/\sigma^2$. A high precision means a narrow distribution (low variance) centered on μ.[4]

Note that, since this is a pdf, we can have $p(x) > 1$. To see this, consider evaluating the density at its center, $x = \mu$. We have $\mathcal{N}(\mu|\mu, \sigma^2) = (\sigma\sqrt{2\pi})^{-1}e^0$, so if $\sigma < 1/\sqrt{2\pi}$, we have $p(x) > 1$.

The cumulative distribution function or cdf of the Gaussian is defined as

$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu, \sigma^2) dz \tag{2.44}$$

See Figure 2.3(a) for a plot of this cdf when $\mu = 0$, $\sigma^2 = 1$. This integral has no closed form expression, but is built in to most software packages. In particular, we can compute it in terms of the **error function** (**erf**):

$$\Phi(x; \mu, \sigma) = \frac{1}{2}[1 + \text{erf}(z/\sqrt{2})] \tag{2.45}$$

4. The symbol λ will have many different meanings in this book, in order to be consistent with the rest of the literature. The intended meaning should be clear from context.

where $z = (x - \mu)/\sigma$ and

$$\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{2.46}$$

The Gaussian distribution is the most widely used distribution in statistics. There are several reasons for this. First, it has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean and variance. Second, the central limit theorem (Section 2.6.3) tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or "noise". Third, the Gaussian distribution makes the least number of assumptions (has maximum entropy), subject to the constraint of having a specified mean and variance, as we show in Section 9.2.6; this makes it a good default choice in many cases. Finally, it has a simple mathematical form, which results in easy to implement, but often highly effective, methods, as we will see. See (Jaynes 2003, ch 7) for a more extensive discussion of why Gaussians are so widely used.

2.4.2 Degenerate pdf

In the limiting case where $\sigma^2 \to 0$, the Gaussian becomes an infinitely tall and infinitely thin "spike" centered at μ:

$$\lim_{\sigma^2 \to 0} \mathcal{N}(x|\mu, \sigma^2) = \delta(x - \mu) \tag{2.47}$$

where δ is called a **Dirac delta function**, and is defined as

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases} \tag{2.48}$$

such that

$$\int_{-\infty}^{\infty} \delta(x) dx = 1 \tag{2.49}$$

A useful property of delta functions is the **sifting property**, which selects out a single term from a sum or integral:

$$\int_{-\infty}^{\infty} f(x)\delta(x - \mu) dx = f(\mu) \tag{2.50}$$

since the integrand is only non-zero if $x - \mu = 0$.

2.4.3 The Student's t distribution

One problem with the Gaussian distribution is that it is sensitive to outliers, since the log-probability only decays quadratically with distance from the center. A more robust distribution is the **Student's t distribution**, which we shall call the Student distribution for short (although

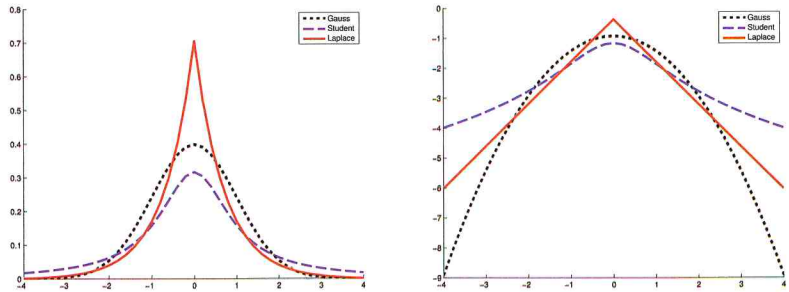


Figure 2.7 (a) The pdf's for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\text{Lap}(0,1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$. (b) Log of these pdf's. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal. Figure generated by `studentLaplacePdfPlot`.

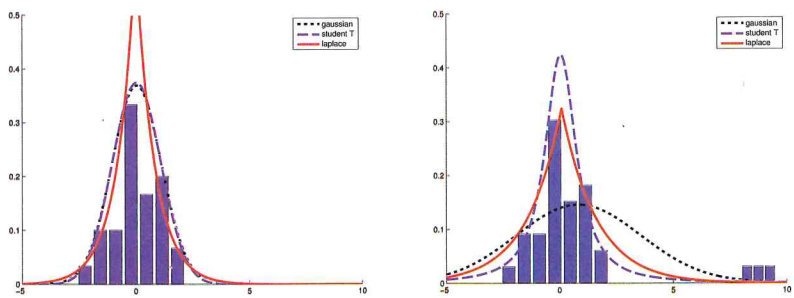


Figure 2.8 Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006). Figure generated by `robustDemo`.

it is more commonly called the t-distribution).[5] Its pdf is as follows:

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\frac{\nu+1}{2})} \tag{2.51}$$

5. This distribution has a colourful etymology. It was first published in 1908 by William Sealy Gosset, who worked at the Guinness brewery in Dublin, Ireland. Since his employer would not allow him to use his own name, he called it the "Student" distribution. The origin of the term t seems to have arisen in the context of tables of the Student distribution, used by Fisher when developing the basis of classical statistical inference. See `http://jeff560.tripod.com/s.html` for more historical details.

where μ is the mean, $\sigma^2 > 0$ is the scale parameter, and $\nu > 0$ is called the **degrees of freedom**. See Figure 2.7 for some plots. For later reference, we note that the distribution has the following properties:

$$\text{mean} = \mu, \text{mode} = \mu, \text{ var} = \frac{\nu\sigma^2}{(\nu - 2)} \tag{2.52}$$

The variance is only defined if $\nu > 2$. The mean is only defined if $\nu > 1$.

As an illustration of the robustness of the Student distribution, consider Figure 2.8. On the left, we show a Gaussian and a Student fit to some data with no outliers. On the right, we add some outliers. We see that the Gaussian is affected a lot, whereas the Student distribution hardly changes. This is because the Student has heavier tails, at least for small ν (see Figure 2.7).

If $\nu = 1$, this distribution is known as the **Cauchy** or **Lorentz** distribution. This is notable for having such heavy tails that the integral that defines the mean does not converge.

To ensure finite variance, we require $\nu > 2$. It is common to use $\nu = 4$, which gives good performance in a range of problems (Lange et al. 1989). For $\nu \gg 5$, the Student distribution rapidly approaches a Gaussian distribution and loses its robustness properties.

2.4.4 The Laplace distribution

Another distribution with heavy tails is the **Laplace distribution**[6], also known as the **double sided exponential** distribution. This has the following pdf:

$$\text{Lap}(x|\mu, b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x - \mu|}{b}\right) \tag{2.53}$$

Here μ is a location parameter and $b > 0$ is a scale parameter. See Figure 2.7 for a plot. This distribution has the following properties:

$$\text{mean} = \mu, \text{ mode} = \mu, \text{ var} = 2b^2 \tag{2.54}$$

Its robustness to outliers is illustrated in Figure 2.8. It also puts more probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model, as we will see in Section 13.3.

2.4.5 The gamma distribution

The **gamma distribution** is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters, called the shape $a > 0$ and the rate $b > 0$:[7]

$$\text{Ga}(T|\text{shape} = a, \text{rate} = b) \triangleq \frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb} \tag{2.55}$$

6. Pierre-Simon Laplace (1749–1827) was a French mathematician, who played a key role in creating the field of Bayesian statistics.

7. There is an alternative parameterization, where we use the scale parameter instead of the rate: $\text{Ga}_s(T|a, b) \triangleq \text{Ga}(T|a, 1/b)$. This version is the one used by Matlab's `gampdf`, although in this book will use the rate parameterization unless otherwise specified.

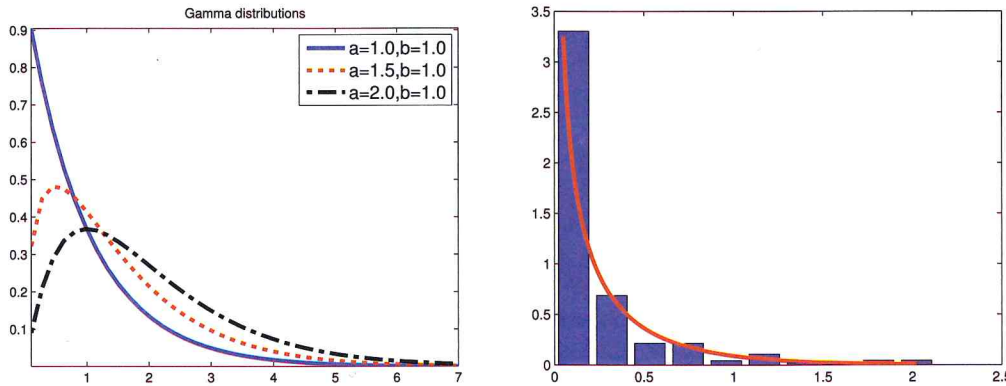


Figure 2.9 (a) Some $\mathrm{Ga}(a, b = 1)$ distributions. If $a \leq 1$, the mode is at 0, otherwise it is > 0. As we increase the rate b, we reduce the horizontal scale, thus squeezing everything leftwards and upwards. Figure generated by `gammaPlotDemo`. (b) An empirical pdf of some rainfall data, with a fitted Gamma distribution superimposed. Figure generated by `gammaRainfallDemo`.

where $\Gamma(a)$ is the gamma function:

$$\Gamma(x) \triangleq \int_0^\infty u^{x-1} e^{-u} du \tag{2.56}$$

See Figure 2.9 for some plots. For later reference, we note that the distribution has the following properties:

$$\text{mean} = \frac{a}{b}, \ \text{mode} = \frac{a-1}{b}, \ \text{var} = \frac{a}{b^2} \tag{2.57}$$

There are several distributions which are just special cases of the Gamma, which we discuss below.

- **Exponential distribution** This is defined by $\mathrm{Expon}(x|\lambda) \triangleq \mathrm{Ga}(x|1, \lambda)$, where λ is the rate parameter. This distribution describes the times between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate λ.
- **Erlang distribution** This is the same as the Gamma distribution where a is an integer. It is common to fix $a = 2$, yielding the one-parameter Erlang distribution, $\mathrm{Erlang}(x|\lambda) = \mathrm{Ga}(x|2, \lambda)$, where λ is the rate parameter.
- **Chi-squared distribution** This is defined by $\chi^2(x|\nu) \triangleq \mathrm{Ga}(x|\frac{\nu}{2}, \frac{1}{2})$. This is the distribution of the sum of squared Gaussian random variables. More precisely, if $Z_i \sim \mathcal{N}(0, 1)$, and $S = \sum_{i=1}^{\nu} Z_i^2$, then $S \sim \chi_\nu^2$.

Another useful result is the following: If $X \sim \mathrm{Ga}(a, b)$, then one can show (Exercise 2.10) that $\frac{1}{X} \sim \mathrm{IG}(a, b)$, where IG is the **inverse gamma** distribution defined by

$$\mathrm{IG}(x|\text{shape} = a, \text{scale} = b) \triangleq \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-b/x} \tag{2.58}$$

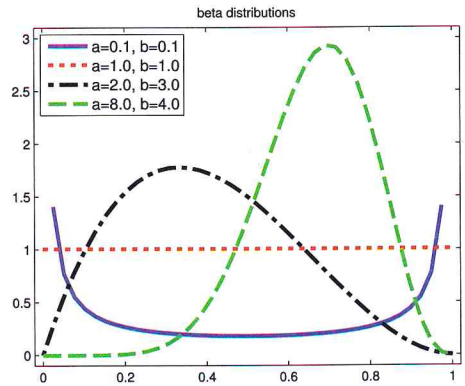


Figure 2.10 Some beta distributions. Figure generated by `betaPlotDemo`.

The distribution has these properties

$$\text{mean} = \frac{b}{a-1}, \quad \text{mode} = \frac{b}{a+1}, \text{var} = \frac{b^2}{(a-1)^2(a-2)}, \tag{2.59}$$

The mean only exists if $a > 1$. The variance only exists if $a > 2$.

We will see applications of these distributions later on.

2.4.6 The beta distribution

The **beta distribution** has support over the interval $[0, 1]$ and is defined as follows:

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \tag{2.60}$$

Here $B(p, q)$ is the beta function,

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.61}$$

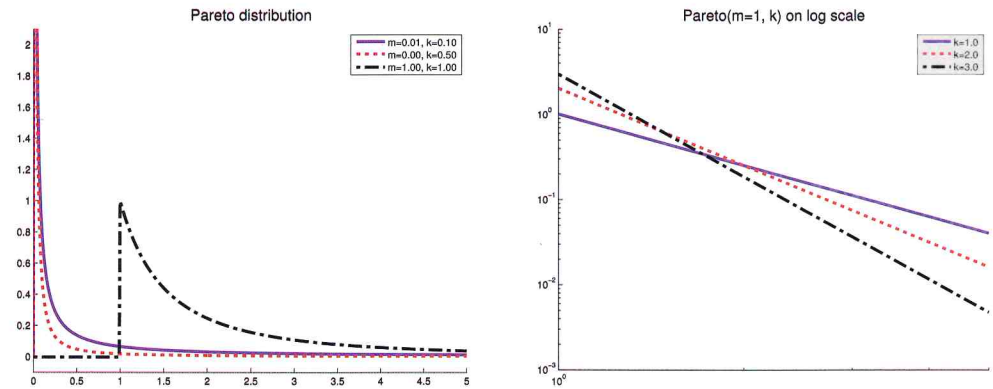
See Figure 2.10 for plots of some beta distributions. We require $a, b > 0$ to ensure the distribution is integrable (i.e., to ensure $B(a, b)$ exists). If $a = b = 1$, we get the uniform distribution. If a and b are both less than 1, we get a bimodal distribution with "spikes" at 0 and 1; if a and b are both greater than 1, the distribution is unimodal. For later reference, we note that the distribution has the following properties (Exercise 2.16):

$$\text{mean} = \frac{a}{a+b}, \quad \text{mode} = \frac{a-1}{a+b-2}, \quad \text{var} = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.62}$$

2.4.7 Pareto distribution

The **Pareto distribution** is defined as follows:

$$\text{Pareto}(x|k, m) = km^k x^{-(k+1)} \mathbb{I}(x \geq m) \tag{2.63}$$



(a) (b)

Figure 2.11 (a) The Pareto distribution $\text{Pareto}(x|m,k)$. (b) The pdf on a log-log scale. Figure generated by `paretoPlot`.

This density asserts that x must be greater than some constant m, but not too much greater, where k controls what is "too much". As $k \to \infty$, the distribution approaches $\delta(x-m)$. See Figure 2.11(a) for some plots. This distribution has the following properties

$$\text{mean} = \frac{km}{k-1} \text{ if } k > 1 \;, \quad \text{mode} = m, \quad \text{var} = \frac{m^2 k}{(k-1)^2(k-2)} \text{ if } k > 2 \tag{2.64}$$

When $m = 0$, the distribution has the form $p(x) = kx^a$, where $a = -(k+1)$. If we plot the distribution on a log-log scale, it forms a straight line, of the form $\log p(x) = a \log x + k$. See Figure 2.11(b) for an illustration; this is known as a **power law**. This is useful for modeling the distribution of quantities that exhibit **long tails**, also called **heavy tails**. For example, it has been observed that the most frequent word in English ("the") occurs approximately twice as often as the second most frequent word ("of"), which occurs twice as often as the fourth most frequent word, etc. If we plot the frequency of words vs their rank, we will get a power law; this is known as **Zipf's law**. Wealth has a similarly skewed distribution, especially in plutocracies such as the USA.[8]

2.5 Joint probability distributions

So far, we have been mostly focusing on modeling univariate probability distributions. In this section, we start our discussion of the more challenging problem of building joint probability distributions on multiple related random variables; this will be a central topic in this book.

A **joint probability distribution** has the form $p(x_1, \ldots, x_D)$ for a set of $D > 1$ variables, and models the (stochastic) relationships between the variables. If all the variables are discrete,

8. In the USA, 400 Americans have more wealth than half of all Americans combined. (Source: `http://www.politifact.com/wisconsin/statements/2011/mar/10/michael-moore/michael-moore-says-400-americans-have-more-wealth-`.) See (Hacker and Pierson 2010) for a political analysis of how such an extreme distribution of income has arisen in a democratic country.

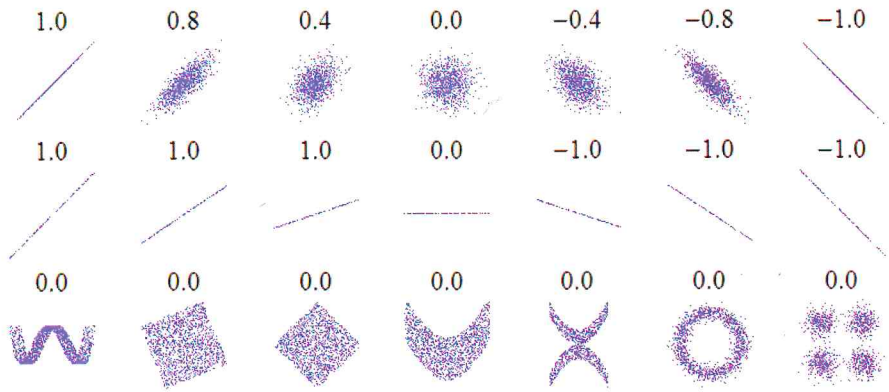


Figure 2.12 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

we can represent the joint distribution as a big multi-dimensional array, with one variable per dimension. However, the number of parameters needed to define such a model is $O(K^D)$, where K is the number of states for each variable.

We can define high dimensional joint distributions using fewer parameters by making conditional independence assumptions, as we explain in Chapter 10. In the case of continuous distributions, an alternative approach is to restrict the form of the pdf to certain functional forms, some of which we will examine below.

2.5.1 Covariance and correlation

The **covariance** between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\text{cov}\left[X, Y\right] \triangleq \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{2.65}$$

If $\mathbf{x}$ is a d-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

$$\text{cov}\left[\mathbf{x}\right] \triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}\left[\mathbf{x}\right])(\mathbf{x} - \mathbb{E}\left[\mathbf{x}\right])^T\right] \tag{2.66}$$

$$= \begin{pmatrix} \text{var}\left[X_1\right] & \text{cov}\left[X_1, X_2\right] & \cdots & \text{cov}\left[X_1, X_d\right] \\ \text{cov}\left[X_2, X_1\right] & \text{var}\left[X_2\right] & \cdots & \text{cov}\left[X_2, X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}\left[X_d, X_1\right] & \text{cov}\left[X_d, X_2\right] & \cdots & \text{var}\left[X_d\right] \end{pmatrix} \tag{2.67}$$

Covariances can be between 0 and infinity. Sometimes it is more convenient to work with a normalized measure, with a finite upper bound. The (Pearson) **correlation coefficient** between

X and Y is defined as

$$\text{corr}\left[X, Y\right] \triangleq \frac{\text{cov}\left[X, Y\right]}{\sqrt{\text{var}\left[X\right]\text{var}\left[Y\right]}} \tag{2.68}$$

A **correlation matrix** has the form

$$\mathbf{R} = \begin{pmatrix} \text{corr}\left[X_1, X_1\right] & \text{corr}\left[X_1, X_2\right] & \cdots & \text{corr}\left[X_1, X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}\left[X_d, X_1\right] & \text{corr}\left[X_d, X_2\right] & \cdots & \text{corr}\left[X_d, X_d\right] \end{pmatrix} \tag{2.69}$$

One can show (Exercise 4.3) that $-1 \leq \text{corr}\left[X, Y\right] \leq 1$. Hence in a correlation matrix, each entry on the diagonal is 1, and the other entries are between -1 and 1.

One can also show that $\text{corr}\left[X, Y\right] = 1$ if and only if $Y = aX + b$ for some parameters a and b, i.e., if there is a *linear* relationship between X and Y (see Exercise 4.4). Intuitively one might expect the correlation coefficient to be related to the slope of the regression line, i.e., the coefficient a in the expression $Y = aX + b$. However, as we show in Equation 7.99 later, the regression coefficient is in fact given by $a = \text{cov}\left[X, Y\right] / \text{var}\left[X\right]$. A better way to think of the correlation coefficient is as a degree of linearity: see Figure 2.12.

If X and Y are independent, meaning $p(X, Y) = p(X)p(Y)$ (see Section 2.2.4), then $\text{cov}\left[X, Y\right] = 0$, and hence $\text{corr}\left[X, Y\right] = 0$ so they are uncorrelated. However, the converse is not true: *uncorrelated does not imply independent.* For example, let $X \sim U(-1, 1)$ and $Y = X^2$. Clearly Y is dependent on X (in fact, Y is uniquely determined by X), yet one can show (Exercise 4.1) that $\text{corr}\left[X, Y\right] = 0$. Some striking examples of this fact are shown in Figure 2.12. This shows several data sets where there is clear dependence between X and Y, and yet the correlation coefficient is 0. A more general measure of dependence between random variables is mutual information, discussed in Section 2.8.3. This is only zero if the variables truly are independent.

2.5.2 The multivariate Gaussian

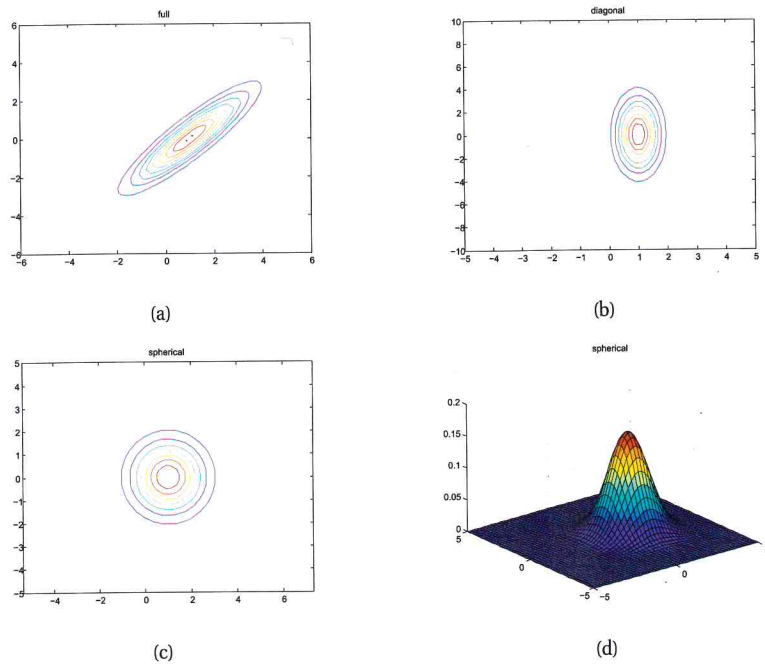
The **multivariate Gaussian** or **multivariate normal** (**MVN**) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] \tag{2.70}$$

where $\boldsymbol{\mu} = \mathbb{E}\left[\mathbf{x}\right] \in \mathbb{R}^D$ is the mean vector, and $\boldsymbol{\Sigma} = \text{cov}\left[\mathbf{x}\right]$ is the $D \times D$ covariance matrix. Sometimes we will work in terms of the **precision matrix** or **concentration matrix** instead. This is just the inverse covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. The normalization constant $(2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}$ just ensures that the pdf integrates to 1 (see Exercise 4.5).

Figure 2.13 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has $D(D+1)/2$ parameters (we divide by 2 since $\boldsymbol{\Sigma}$ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A **spherical** or **isotropic** covariance, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_D$, has one free parameter.



(a) (b) (c) (d)

Figure 2.13 We show the level sets for 2d Gaussians. (a) A full covariance matrix has elliptical contours. (b) A diagonal covariance matrix is an **axis aligned** ellipse. (c) A spherical covariance matrix has a circular shape. (d) Surface plot for the spherical Gaussian in (c). Figure generated by `gaussPlot2Ddemo`.

2.5.3 Multivariate Student t distribution

A more robust alternative to the MVN is the **multivariate Student t** distribution, whose pdf is given by

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Sigma}|^{-1/2}}{\nu^{D/2}\pi^{D/2}} \times \left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \quad (2.71)$$

$$= \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}|\pi\mathbf{V}|^{-1/2} \times \left[1+(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \quad (2.72)$$

where $\boldsymbol{\Sigma}$ is called the scale matrix (since it is not exactly the covariance matrix) and $\mathbf{V}=\nu\boldsymbol{\Sigma}$. This has fatter tails than a Gaussian. The smaller ν is, the fatter the tails. As $\nu\to\infty$, the distribution tends towards a Gaussian. The distribution has the following properties

$$\text{mean} = \boldsymbol{\mu}, \quad \text{mode} = \boldsymbol{\mu}, \quad \text{Cov} = \frac{\nu}{\nu-2}\boldsymbol{\Sigma} \quad (2.73)$$

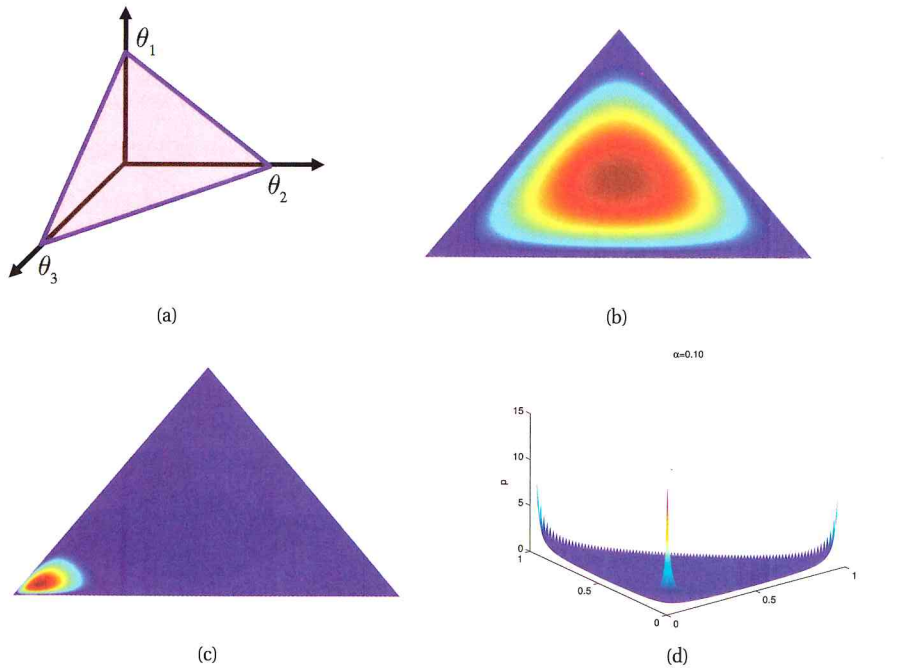


Figure 2.14 (a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{3} \theta_k = 1$. (b) Plot of the Dirichlet density when $\boldsymbol{\alpha} = (2, 2, 2)$. (c) $\boldsymbol{\alpha} = (20, 2, 2)$. Figure generated by `visDirichletGui`, by Jonathan Huang. (d) $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$. (The comb-like structure on the edges is a plotting artifact.) Figure generated by `dirichlet3dPlot`.

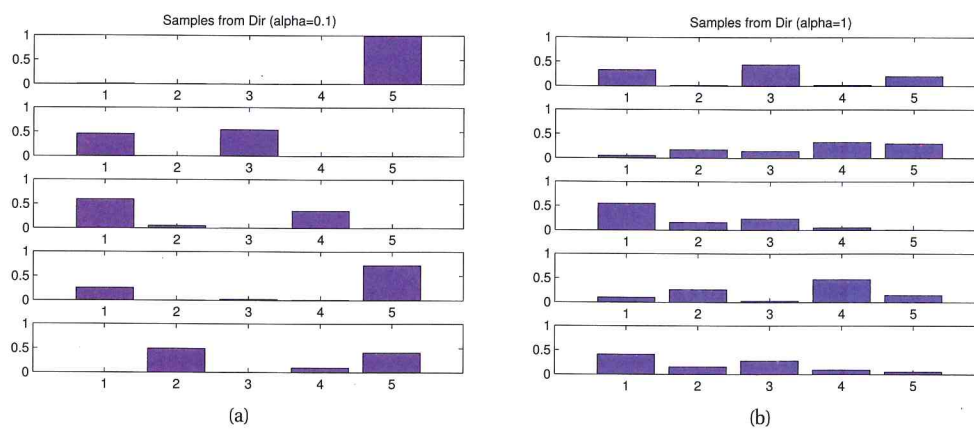


Figure 2.15 Samples from a 5-dimensional symmetric Dirichlet distribution for different parameter values. (a) $\boldsymbol{\alpha} = (0.1, \ldots, 0.1)$. This results in very sparse distributions, with many 0s. (b) $\boldsymbol{\alpha} = (1, \ldots, 1)$. This results in more uniform (and dense) distributions. Figure generated by `dirichletHistogramDemo`.

2.5.4 Dirichlet distribution

A multivariate generalization of the beta distribution is the **Dirichlet**[9] distribution, which has support over the **probability simplex**, defined by

$$S_K = \{\mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1\} \tag{2.74}$$

The pdf is defined as follows:

$$\text{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K) \tag{2.75}$$

where $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables:

$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \tag{2.76}$$

where $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$.

Figure 2.14 shows some plots of the Dirichlet when $K = 3$, and Figure 2.15 for some sampled probability vectors. We see that $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ controls the strength of the distribution (how peaked it is), and the α_k control where the peak occurs. For example, $\text{Dir}(1, 1, 1)$ is a uniform distribution, $\text{Dir}(2, 2, 2)$ is a broad distribution centered at $(1/3, 1/3, 1/3)$, and $\text{Dir}(20, 20, 20)$ is a narrow distribution centered at $(1/3, 1/3, 1/3)$. If $\alpha_k < 1$ for all k, we get "spikes" at the corners of the simplex.

For future reference, the distribution has these properties

$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \; \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \; \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.77}$$

where $\alpha_0 = \sum_k \alpha_k$. Often we use a symmetric Dirichlet prior of the form $\alpha_k = \alpha/K$. In this case, the mean becomes $1/K$, and the variance becomes $\text{var}[x_k] = \frac{K-1}{K^2(\alpha+1)}$. So increasing α increases the precision (decreases the variance) of the distribution.

2.6 Transformations of random variables

If $\mathbf{x} \sim p()$ is some random variable, and $\mathbf{y} = f(\mathbf{x})$, what is the distribution of $\mathbf{y}$? This is the question we address in this section.

2.6.1 Linear transformations

Suppose $f()$ is a linear function:

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{2.78}$$

9. Johann Dirichlet was a German mathematician, 1805–1859.

In this case, we can easily derive the mean and covariance of $\mathbf{y}$ as follows. First, for the mean, we have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{2.79}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. This is called the **linearity of expectation**. If $f()$ is a scalar-valued function, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$, the corresponding result is

$$\mathbb{E}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\mu} + b \tag{2.80}$$

For the covariance, we have

$$\text{cov}[\mathbf{y}] = \text{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \tag{2.81}$$

where $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$. We leave the proof of this as an exercise. If $f()$ is scalar valued, the result becomes

$$\text{var}[y] = \text{var}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} \tag{2.82}$$

We will use both of these results extensively in later chapters. Note, however, that the mean and covariance only completely define the distribution of $\mathbf{y}$ if $\mathbf{x}$ is Gaussian. In general we must use the techniques described below to derive the full distribution of $\mathbf{y}$, as opposed to just its first two moments.

2.6.2 General transformations

If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the x's such that $f(x) = y$:

$$p_y(y) = \sum_{x:f(x)=y} p_x(x) \tag{2.83}$$

For example, if $f(X) = 1$ if X is even and $f(X) = 0$ otherwise, and $p_x(X)$ is uniform on the set $\{1, \ldots, 10\}$, then $p_y(1) = \sum_{x\in\{2,4,6,8,10\}} p_x(x) = 0.5$, and $p_y(0) = 0.5$ similarly. Note that in this example, f is a many-to-one function.

If X is continuous, we cannot use Equation 2.83 since $p_x(x)$ is a density, not a pmf, and we cannot sum up densities. Instead, we work with cdf's, and write

$$P_y(y) \triangleq P(Y \leq y) = P(f(X) \leq y) = P(X \in \{x|f(x) \leq y\}) \tag{2.84}$$

We can derive the pdf of y by differentiating the cdf.

In the case of monotonic and hence invertible functions, we can write

$$P_y(y) = P(f(X) \leq y) = P(X \leq f^{-1}(y)) = P_x(f^{-1}(y)) \tag{2.85}$$

Taking derivatives we get

$$p_y(y) \triangleq \frac{d}{dy}P_y(y) = \frac{d}{dy}P_x(f^{-1}(y)) = \frac{dx}{dy}\frac{d}{dx}P_x(x) = \frac{dx}{dy}p_x(x) \tag{2.86}$$

where $x = f^{-1}(y)$. We can think of dx as a measure of volume in the x-space; similarly dy measures volume in y space. Thus $\frac{dx}{dy}$ measures the change in volume. Since the sign of this change is not important, we take the absolute value to get the general expression:

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| \tag{2.87}$$

This is called **change of variables** formula. We can understand this result more intuitively as follows. Observations falling in the range $(x, x+\delta x)$ will get transformed into $(y, y+\delta y)$, where $p_x(x)\delta x \approx p_y(y)\delta_y$. Hence $p_y(y) \approx p_x(x)|\frac{\delta x}{\delta y}|$. For example, suppose $X \sim U(-1, 1)$, and $Y = X^2$. Then $p_y(y) = \frac{1}{2}y^{-\frac{1}{2}}$. See also Exercise 2.10.

2.6.2.1 Multivariate change of variables *

We can extend the previous results to multivariate distributions as follows. Let f be a function that maps $\mathbb{R}^n$ to $\mathbb{R}^n$, and let $\mathbf{y} = f(\mathbf{x})$. Then its **Jacobian matrix** $\mathbf{J}$ is given by

$$\mathbf{J}_{\mathbf{x}\to\mathbf{y}} \triangleq \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \tag{2.88}$$

$|\det \mathbf{J}|$ measures how much a unit cube changes in volume when we apply f.

If f is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\mathbf{y} \to \mathbf{x}$:

$$p_y(\mathbf{y}) = p_x(\mathbf{x})\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right| = p_x(\mathbf{x})|\det \mathbf{J}_{\mathbf{y}\to\mathbf{x}}| \tag{2.89}$$

In Exercise 4.5 you will use this formula to derive the normalization constant for a multivariate Gaussian.

As a simple example, consider transforming a density from Cartesian coordinates $\mathbf{x} = (x_1, x_2)$ to polar coordinates $\mathbf{y} = (r, \theta)$, where $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$. Then

$$\mathbf{J}_{\mathbf{y}\to\mathbf{x}} = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \tag{2.90}$$

and

$$|\det \mathbf{J}| = |r\cos^2\theta + r\sin^2\theta| = |r| \tag{2.91}$$

Hence

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})|\det \mathbf{J}| \tag{2.92}$$

$$p_{r,\theta}(r, \theta) = p_{x_1,x_2}(x_1, x_2)r = p_{x_1,x_2}(r\cos\theta, r\sin\theta)r \tag{2.93}$$

To see this geometrically, notice that the area of the shaded patch in Figure 2.16 is given by

$$P(r \le R \le r + dr, \theta \le \Theta \le \theta + d\theta) = p_{r,\theta}(r, \theta)drd\theta \tag{2.94}$$

In the limit, this is equal to the density at the center of the patch, $p(r, \theta)$, times the size of the patch, $r\ dr\ d\theta$. Hence

$$p_{r,\theta}(r, \theta)drd\theta = p_{x_1,x_2}(r\cos\theta, r\sin\theta)r\ dr\ d\theta \tag{2.95}$$

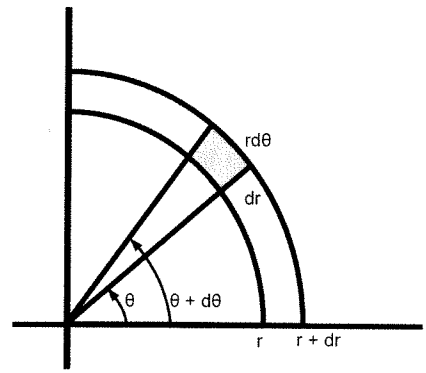


Figure 2.16 Change of variables from polar to Cartesian. The area of the shaded patch is $r\ dr\ d\theta$. Based on (Rice 1995) Figure 3.16.

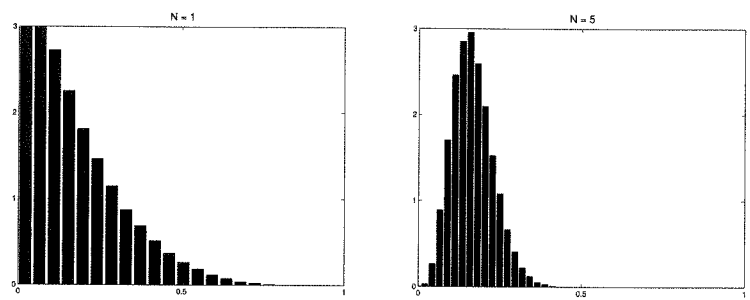


Figure 2.17 The central limit theorem in pictures. We plot a histogram of $\frac{1}{N}\sum_{i=1}^{N} x_{ij}$, where $x_{ij} \sim \text{Beta}(1,5)$, for $j = 1:10000$. As $N \to \infty$, the distribution tends towards a Gaussian. (a) $N = 1$. (b) $N = 5$. Based on Figure 2.6 of (Bishop 2006). Figure generated by `centralLimitDemo`.

2.6.3 Central limit theorem

Now consider N random variables with pdf's (not necessarily Gaussian) $p(x_i)$, each with mean μ and variance σ^2. We assume each variable is **independent and identically distributed** or **iid** for short. Let $S_N = \sum_{i=1}^{N} X_i$ be the sum of the rv's. This is a simple but widely used transformation of rv's. One can show that, as N increases, the distribution of this sum approaches

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right) \tag{2.96}$$

Hence the distribution of the quantity

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \tag{2.97}$$

converges to the standard normal, where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean. This is called the **central limit theorem**. See e.g., (Jaynes 2003, p222) or (Rice 1995, p169) for a proof.

In Figure 2.17 we give an example in which we compute the mean of rv's drawn from a beta distribution. We see that the sampling distribution of the mean value rapidly converges to a Gaussian distribution.

2.7 Monte Carlo approximation

In general, computing the distribution of a function of an rv using the change of variables formula can be difficult. One simple but powerful alternative is as follows. First we generate S samples from the distribution, call them $x_1, \ldots, x_S$. (There are many ways to generate such samples; one popular method, for high dimensional distributions, is called Markov chain Monte Carlo or MCMC; this will be explained in Chapter 24.) Given the samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$. This is called a **Monte Carlo** approximation, named after a city in Europe known for its plush gambling casinos. Monte Carlo techniques were first developed in the area of statistical physics — in particular, during development of the atomic bomb — but are now widely used in statistics and machine learning as well.

We can use Monte Carlo to approximate the expected value of any function of a random variable. We simply draw samples, and then compute the arithmetic mean of the function applied to the samples. This can be written as follows:

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) \tag{2.98}$$

where $x_s \sim p(X)$. This is called Monte Carlo integration, and has the advantage over numerical integration (which is based on evaluating the function at a fixed grid of points) that the function is only evaluated in places where there is non-negligible probability.

By varying the function $f()$, we can approximate many quantities of interest, such as

- $\overline{x} = \frac{1}{S}\sum_{s=1}^{S} x_s \to \mathbb{E}[X]$
- $\frac{1}{S}\sum_{s=1}^{S}(x_s - \overline{x})^2 \to \text{var}[X]$
- $\frac{1}{S}|\{x_s \leq c\}| \to P(X \leq c)$
- $\text{median}\{x_1, \ldots, x_S\} \to \text{median}(X)$

We give some examples below, and will see many more in later chapters.

2.7.1 Example: change of variables, the MC way

In Section 2.6.2, we discussed how to analytically compute the distribution of a function of a random variable, $y = f(x)$. A much simpler approach is to use a Monte Carlo approximation. For example, suppose $x \sim \text{Unif}(-1, 1)$ and $y = x^2$. We can approximate $p(y)$ by drawing many samples from $p(x)$, squaring them, and computing the resulting empirical distribution. See Figure 2.18 for an illustration. We will use this technique extensively in later chapters. See also Figure 5.2.

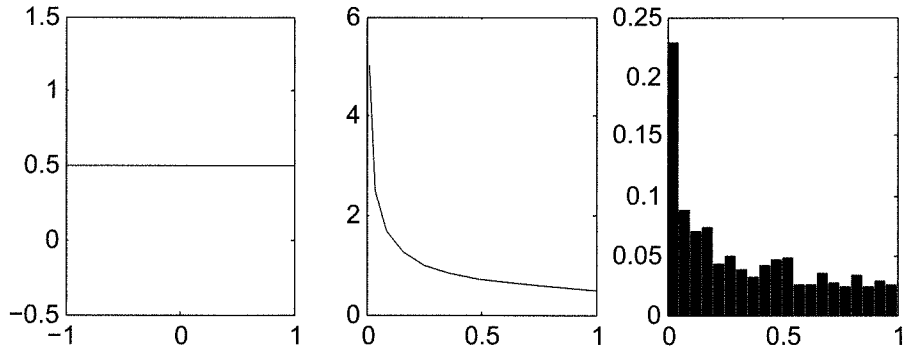


Figure 2.18 Computing the distribution of $y = x^2$, where $p(x)$ is uniform (left). The analytic result is shown in the middle, and the Monte Carlo approximation is shown on the right. Figure generated by `changeOfVarsDemo1d`.

2.7.2 Example: estimating π by Monte Carlo integration

MC approximation can be used for many applications, not just statistical ones. Suppose we want to estimate π. We know that the area of a circle with radius r is πr^2, but it is also equal to the following definite integral:

$$I = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy \tag{2.99}$$

Hence $\pi = I/(r^2)$. Let us approximate this by Monte Carlo integration. Let $f(x, y) = \mathbb{I}(x^2 + y^2 \leq r^2)$ be an indicator function that is 1 for points inside the circle, and 0 outside, and let $p(x)$ and $p(y)$ be uniform distributions on $[-r, r]$, so $p(x) = p(y) = 1/(2r)$. Then

$$I = (2r)(2r) \int \int f(x, y) p(x) p(y) dx dy \tag{2.100}$$

$$= 4r^2 \int \int f(x, y) p(x) p(y) dx dy \tag{2.101}$$

$$\approx 4r^2 \frac{1}{S} \sum_{s=1}^{S} f(x_s, y_s) \tag{2.102}$$

We find $\hat{\pi} = 3.1416$ with standard error 0.09 (see Section 2.7.3 for a discussion of standard errors). We can plot the points that are accepted or rejected as in Figure 2.19.

2.7.3 Accuracy of Monte Carlo approximation

The accuracy of an MC approximation increases with sample size. This is illustrated in Figure 2.20, On the top line, we plot a histogram of samples from a Gaussian distribution. On the bottom line, we plot a smoothed version of these samples, created using a kernel density

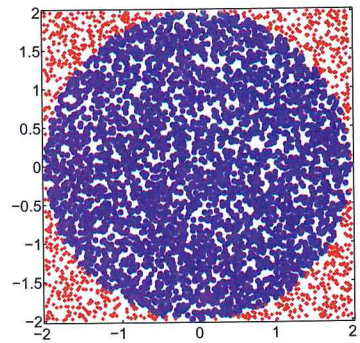


Figure 2.19 Estimating π by Monte Carlo integration. Blue points are inside the circle, red crosses are outside. Figure generated by `mcEstimatePi`.

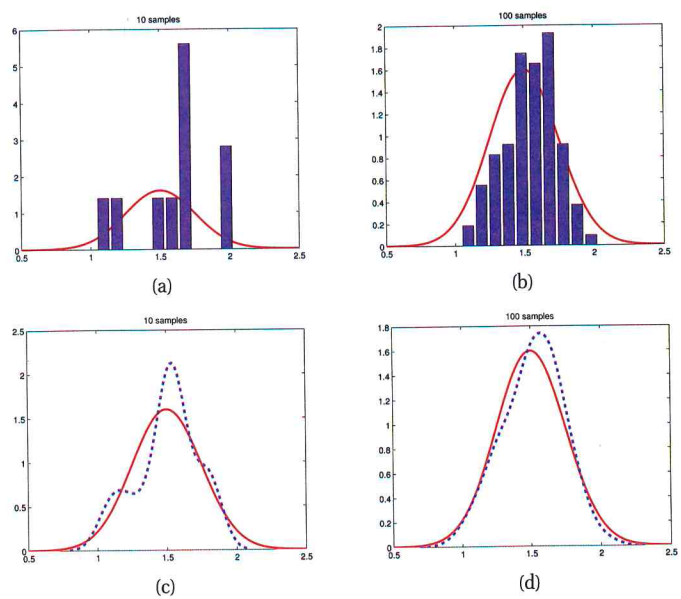


Figure 2.20 10 and 100 samples from a Gaussian distribution, $\mathcal{N}(\mu = 1.5, \sigma^2 = 0.25)$. Solid red line is true pdf. Top line: histogram of samples. Bottom line: kernel density estimate derived from samples in dotted blue, solid red line is true pdf. Based on Figure 4.1 of (Hoff 2009). Figure generated by `mcAccuracyDemo`.

estimate (Section 14.7.2). This smoothed distribution is then evaluated on a dense grid of points and plotted. Note that this smoothing is just for the purposes of plotting, it is not used for the Monte Carlo estimate itself.

If we denote the exact mean by $\mu = \mathbb{E}[f(X)]$, and the MC approximation by $\hat{\mu}$, one can show that, with independent samples,

$$(\hat{\mu} - \mu) \rightarrow \mathcal{N}(0, \frac{\sigma^2}{S}) \tag{2.103}$$

where

$$\sigma^2 = \text{var}[f(X)] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}[f(X)]^2 \tag{2.104}$$

This is a consequence of the central-limit theorem. Of course, σ^2 is unknown in the above expression, but it can also be estimated by MC:

$$\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}(f(x_s) - \hat{\mu})^2 \tag{2.105}$$

Then we have

$$P\left\{\mu - 1.96\frac{\hat{\sigma}}{\sqrt{S}} \leq \hat{\mu} \leq \mu + 1.96\frac{\hat{\sigma}}{\sqrt{S}}\right\} \approx 0.95 \tag{2.106}$$

The term $\sqrt{\frac{\hat{\sigma}^2}{S}}$ is called the (numerical or empirical) **standard error**, and is an estimate of our uncertainty about our estimate of μ. (See Section 6.2 for more discussion on standard errors.)

If we want to report an answer which is accurate to within $\pm\epsilon$ with probability at least 95%, we need to use a number of samples S which satisfies $1.96\sqrt{\hat{\sigma}^2/S} \leq \epsilon$. We can approximate the 1.96 factor by 2, yielding $S \geq \frac{4\hat{\sigma}^2}{\epsilon^2}$.

2.8 Information theory

Information theory is concerned with representing data in a compact fashion (a task known as **data compression** or **source coding**), as well as with transmitting and storing it in a way that is robust to errors (a task known as **error correction** or **channel coding**). At first, this seems far removed from the concerns of probability theory and machine learning, but in fact there is an intimate connection. To see this, note that compactly representing data requires allocating short codewords to highly probable bit strings, and reserving longer codewords to less probable bit strings. This is similar to the situation in natural language, where common words (such as "a", "the", "and") are generally much shorter than rare words. Also, decoding messages sent over noisy channels requires having a good probability model of the kinds of messages that people tend to send. In both cases, we need a model that can predict which kinds of data are likely and which unlikely, which is also a central problem in machine learning (see (MacKay 2003) for more details on the connection between information theory and machine learning).

Obviously we cannot go into the details of information theory here (see e.g., (Cover and Thomas 2006) if you are interested to learn more). However, we will introduce a few basic concepts that we will need later in the book.

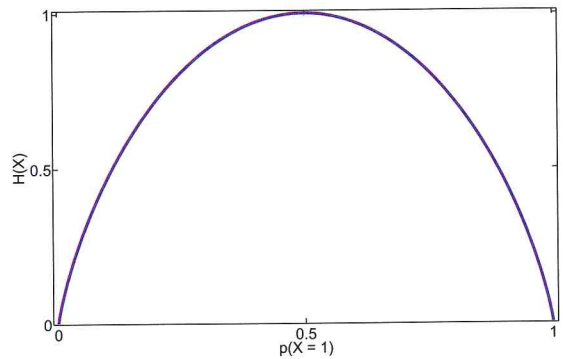


Figure 2.21 Entropy of a Bernoulli random variable as a function of θ. The maximum entropy is $\log_2 2 = 1$. Figure generated by `bernoulliEntropyFig`.

2.8.1 Entropy

The **entropy** of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k) \tag{2.107}$$

Usually we use log base 2, in which case the units are called **bits** (short for binary digits). If we use log base e, the units are called **nats**. For example, if $X \in \{1, \ldots, 5\}$ with histogram distribution $p = [0.25, 0.25, 0.2, 0.15, 0.15]$, we find $H = 2.2855$. The discrete distribution with maximum entropy is the uniform distribution (see Section 9.2.6 for a proof). Hence for a K-ary random variable, the entropy is maximized if $p(x = k) = 1/K$; in this case, $\mathbb{H}(X) = \log_2 K$. Conversely, the distribution with minimum entropy (which is zero) is any delta-function that puts all its mass on one state. Such a distribution has no uncertainty. In Figure 2.5(b), where we plotted a DNA sequence logo, the height of each bar is defined to be $2 - H$, where H is the entropy of that distribution, and 2 is the maximum possible entropy. Thus a bar of height 0 corresponds to a uniform distribution, whereas a bar of height 2 corresponds to a deterministic distribution.

For the special case of binary random variables, $X \in \{0, 1\}$, we can write $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$. Hence the entropy becomes

$$\mathbb{H}(X) = -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \tag{2.108}$$

$$= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)] \tag{2.109}$$

This is called the **binary entropy function**, and is also written $\mathbb{H}(\theta)$. We plot this in Figure 2.21. We see that the maximum value of 1 occurs when the distribution is uniform, $\theta = 0.5$.

2.8.2 KL divergence

One way to measure the dissimilarity of two probability distributions, p and q, is known as the **Kullback-Leibler divergence** (**KL divergence**) or **relative entropy**. This is defined as follows:

$$\mathbb{KL}\left(p||q\right) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \tag{2.110}$$

where the sum gets replaced by an integral for pdfs.[10] We can rewrite this as

$$\mathbb{KL}\left(p||q\right) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H}\left(p\right) + \mathbb{H}\left(p, q\right) \tag{2.111}$$

where $\mathbb{H}\left(p, q\right)$ is called the **cross entropy**,

$$\mathbb{H}\left(p, q\right) \triangleq -\sum_k p_k \log q_k \tag{2.112}$$

One can show (Cover and Thomas 2006) that the cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook. Hence the "regular" entropy $\mathbb{H}\left(p\right) = \mathbb{H}\left(p, p\right)$, defined in Section 2.8.1, is the expected number of bits if we use the true model, so the KL divergence is the difference between these. In other words, the KL divergence is the average number of *extra* bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

The "extra number of bits" interpretation should make it clear that $\mathbb{KL}\left(p||q\right) \geq 0$, and that the KL is only equal to zero iff $q = p$. We now give a proof of this important result.

Theorem 2.8.1. (**Information inequality**) $\mathbb{KL}\left(p||q\right) \geq 0$ *with equality iff* $p = q$.

Proof. To prove the theorem, we need to use **Jensen's inequality**. This states that, for any convex function f, we have that

$$f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} \lambda_i f(\mathbf{x}_i) \tag{2.113}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. This is clearly true for $n = 2$ (by definition of convexity), and can be proved by induction for $n > 2$.

Let us now prove the main theorem, following (Cover and Thomas 2006, p28). Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Then

$$-\mathbb{KL}\left(p||q\right) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{2.114}$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} = \log \sum_{x \in A} q(x) \tag{2.115}$$

$$\leq \log \sum_{x \in \mathcal{X}} q(x) = \log 1 = 0 \tag{2.116}$$

10. The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the **Jensen-Shannon divergence**, defined as $JS(p_1, p_2) = 0.5\mathbb{KL}\left(p_1||q\right) + 0.5\mathbb{KL}\left(p_2||q\right)$, where $q = 0.5p_1 + 0.5p_2$.

where the first inequality follows from Jensen's. Since $\log(x)$ is a strictly concave function, we have equality in Equation 2.115 iff $p(x) = cq(x)$ for some c. We have equality in Equation 2.116 iff $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies $c = 1$. Hence $\mathbb{KL}(p||q) = 0$ iff $p(x) = q(x)$ for all x. □

One important consequence of this result is that the *discrete distribution with the maximum entropy is the uniform distribution.* More precisely, $\mathbb{H}(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ is the number of states for X, with equality iff $p(x)$ is uniform. To see this, let $u(x) = 1/|\mathcal{X}|$. Then

$$0 \leq \mathbb{KL}(p||u) = \sum_x p(x) \log \frac{p(x)}{u(x)} \tag{2.117}$$

$$= \sum_x p(x) \log p(x) - \sum_x p(x) \log u(x) = -\mathbb{H}(X) + \log |\mathcal{X}| \tag{2.118}$$

This is a formulation of Laplace's **principle of insufficient reason**, which argues in favor of using uniform distributions when there are no other reasons to favor one distribution over another. See Section 9.2.6 for a discussion of how to create distributions that satisfy certain constraints, but otherwise are as least-commital as possible. (For example, the Gaussian satisfies first and second moment constraints, but otherwise has maximum entropy.)

2.8.3 Mutual information

Consider two random variables, X and Y. Suppose we want to know how much knowing one variable tells us about the other. We could compute the correlation coefficient, but this is only defined for real-valued random variables, and furthermore, this is a very limited measure of dependence, as we saw in Figure 2.12. A more general approach is to determine how similar the joint distribution $p(X, Y)$ is to the factored distribution $p(X)p(Y)$. This is called the **mutual information** or **MI**, and is defined as follows:

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}(p(X,Y)||p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.119}$$

We have $\mathbb{I}(X;Y) \geq 0$ with equality iff $p(X,Y) = p(X)p(Y)$. That is, the MI is zero iff the variables are independent.

To gain insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies. One can show (Exercise 2.12) that the above expression is equivalent to the following:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \tag{2.120}$$

where $\mathbb{H}(Y|X)$ is the **conditional entropy**, defined as $\mathbb{H}(Y|X) = \sum_x p(x)\mathbb{H}(Y|X = x)$. Thus we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X. We will encounter several applications of MI later in the book. See also Exercises 2.13 and 2.14 for the connection between MI and correlation coefficients.

A quantity which is closely related to MI is the **pointwise mutual information** or PMI. For two events (not random variables) x and y, this is defined as

$$\text{PMI}(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.121}$$

This measures the discrepancy between these events occuring together compared to what would be expected by chance. Clearly the MI of X and Y is just the expected value of the PMI. Interestingly, we can rewrite the PMI as follows:

$$\text{PMI}(x, y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.122}$$

This is the amount we learn from updating the prior $p(x)$ into the posterior $p(x|y)$, or equivalently, updating the prior $p(y)$ into the posterior $p(y|x)$.

2.8.3.1 Mutual information for continuous random variables *

The above formula for MI is defined for discrete random variables. For continuous random variables, it is common to first **discretize** or **quantize** them, by dividing the ranges of each variable into bins, and computing how many values fall in each histogram bin (Scott 1979). We can then easily compute the MI using the formula above (see `mutualInfoAllPairsMixed` for some code, and `miMixedDemo` for a demo).

Unfortunately, the number of bins used, and the location of the bin boundaries, can have a significant effect on the results. One way around this is to try to estimate the MI directly, without first performing density estimation (Learned-Miller 2004). Another approach is to try many different bin sizes and locations, and to compute the maximum MI achieved. This statistic, appropriately normalized, is known as the **maximal information coefficient** (MIC) (Reshef et al. 2011). More precisely, define

$$m(x, y) = \frac{\max_{G \in \mathcal{G}(x,y)} \mathbb{I}\left(X(G); Y(G)\right)}{\log \min(x, y)} \tag{2.123}$$

where $\mathcal{G}(x, y)$ is the set of 2d grids of size $x \times y$, and $X(G)$, $Y(G)$ represents a discretization of the variables onto this grid. (The maximization over bin locations can be performed efficiently using dynamic programming (Reshef et al. 2011).) Now define the MIC as

$$\text{MIC} \triangleq \max_{x,y:xy<B} m(x, y) \tag{2.124}$$

where B is some sample-size dependent bound on the number of bins we can use and still reliably estimate the distribution ((Reshef et al. 2011) suggest $B = N^{0.6}$). It can be shown that the MIC lies in the range $[0, 1]$, where 0 represents no relationship between the variables, and 1 represents a noise-free relationship of any form, not just linear.

Figure 2.22 gives an example of this statistic in action. The data consists of 357 variables measuring a variety of social, economic, health and political indicators, collected by the World Health Organization (WHO). On the left of the figure, we see the correlation coefficient (CC) plotted against the MIC for all 63,566 variable pairs. On the right of the figure, we see scatter plots for particular pairs of variables, which we now discuss:

- The point marked C has a low CC and a low MIC. The corresponding scatter plot makes it clear that there is no relationship between these two variables (percentage of lives lost to injury and density of dentists in the population).
- The points marked D and H have high CC (in absolute value) and high MIC, because they represent nearly linear relationships.

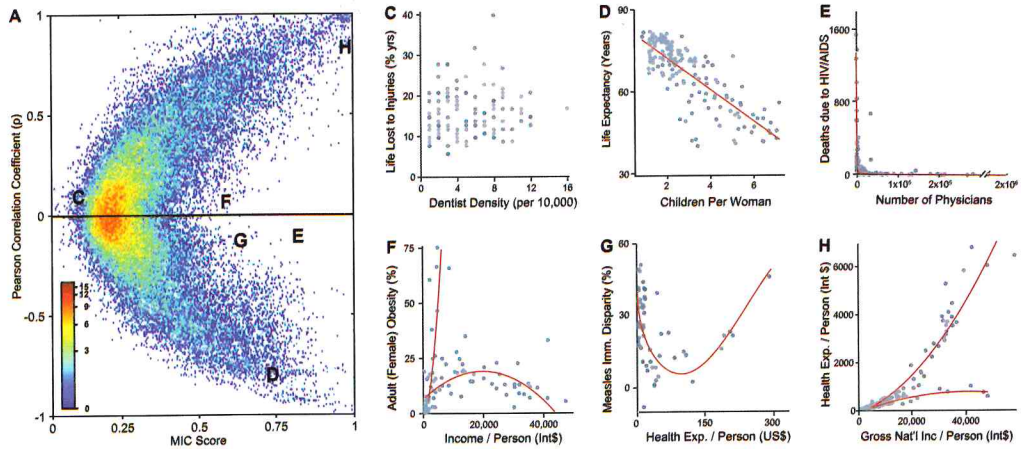


Figure 2.22 Left: Correlation coefficient vs maximal information criterion (MIC) for all pairwise relationships in the WHO data. Right: scatter plots of certain pairs of variables. The red lines are non-parametric smoothing regressions (Section 15.4.6) fit separately to each trend. Source: Figure 4 of (Reshef et al. 2011) . Used with kind permission of David Reshef and the American Association for the Advancement of Science.

- The points marked E, F, and G have low CC but high MIC. This is because they correspond to non-linear (and sometimes, as in the case of E and F, non-functional, i.e., one-to-many) relationships between the variables.

In summary, we see that statistics (such as MIC) based on mutual information can be used to discover interesting relationships between variables in a way that simpler measures, such as correlation coefficients, cannot. For this reason, the MIC has been called "a correlation for the 21st century" (Speed 2011).

Exercises

Exercise 2.1 Probabilities are sensitive to the form of the question that was used to generate the answer

(Source: Minka.) My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities—two boys or two girls—have probabilities 1/4 and 1/4.

a. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?
b. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

Exercise 2.2 Legal reasoning

(Source: Peter Lee.) Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

a. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he is guilty". This is known as the **prosecutor's fallacy**. What is wrong with this argument?

b. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance." This is known as the **defender's fallacy**. What is wrong with this argument?

Exercise 2.3 Variance of a sum

Show that the variance of a sum is $\text{var}[X+Y] = \text{var}[X] + \text{var}[Y] + 2\text{cov}[X,Y]$, where $\text{cov}[X,Y]$ is the covariance between X and Y.

Exercise 2.4 Bayes rule for medical diagnosis

(Source: Koller.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

Exercise 2.5 The Monty Hall problem

(Source: Mackay.) On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors. Hint: use Bayes rule.

Exercise 2.6 Conditional independence

(Source: Koller.)

a. Let $H \in \{1, \ldots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = (P(H=1|e_1, e_2), \ldots, P(H=K|e_1, e_2))$$

Which of the following sets of numbers are sufficient for the calculation?

i. $P(e_1, e_2)$, $P(H)$, $P(e_1|H)$, $P(e_2|H)$
ii. $P(e_1, e_2)$, $P(H)$, $P(e_1, e_2|H)$
iii. $P(e_1|H)$, $P(e_2|H)$, $P(H)$

b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficient now?

Show your calculations as well as giving the final result. Hint: use Bayes rule.

Exercise 2.7 Pairwise independence does not imply mutual independence

We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2.125}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2.126}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \quad \forall S \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{2.127}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{2.128}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

Exercise 2.8 Conditional independence iff joint factorizes

In the text we said $X \perp Y|Z$ iff

$$p(x, y|z) = p(x|z)p(y|z) \tag{2.129}$$

for all x, y, z such that $p(z) > 0$. Now prove the following alternative definition: $X \perp Y|Z$ iff there exist functions g and h such that

$$p(x, y|z) = g(x, z)h(y, z) \tag{2.130}$$

for all x, y, z such that $p(z) > 0$.

Exercise 2.9 Conditional independence

(Source: Koller.) Are the following properties true? Prove or disprove. Note that we are not restricting attention to distributions that can be represented by a graphical model.

a. True or false? $(X \perp W|Z, Y) \wedge (X \perp Y|Z) \Rightarrow (X \perp Y, W|Z)$
b. True or false? $(X \perp Y|Z) \wedge (X \perp Y|W) \Rightarrow (X \perp Y|Z, W)$

Exercise 2.10 Deriving the inverse gamma density

Let $X \sim \text{Ga}(a, b)$, i.e.

$$\text{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb} \tag{2.131}$$

Let $Y = 1/X$. Show that $Y \sim \text{IG}(a, b)$, i.e.,

$$\text{IG}(x|\text{shape} = a, \text{scale} = b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-b/x} \tag{2.132}$$

Hint: use the change of variables formula.

Exercise 2.11 Normalization constant for a 1D Gaussian

The normalization constant for a zero-mean Gaussian is given by

$$Z = \int_a^b \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \tag{2.133}$$

where $a = -\infty$ and $b = \infty$. To compute this, consider its square

$$Z^2 = \int_a^b \int_a^b \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dxdy \tag{2.134}$$

Let us change variables from cartesian (x, y) to polar (r, θ) using $x = r\cos\theta$ and $y = r\sin\theta$. Since $dxdy = rdrd\theta$, and $cos^2\theta + \sin^2\theta = 1$, we have

$$Z^2 = \int_0^{2\pi} \int_0^{\infty} r\exp\left(-\frac{r^2}{2\sigma^2}\right) drd\theta \tag{2.135}$$

Evaluate this integral and hence show $Z = \sqrt{\sigma^2 2\pi}$. Hint 1: separate the integral into a product of two terms, the first of which (involving $d\theta$) is constant, so is easy. Hint 2: if $u = e^{-r^2/2\sigma^2}$ then $du/dr = -\frac{1}{\sigma^2} re^{-r^2/2\sigma^2}$, so the second integral is also easy (since $\int u'(r)dr = u(r)$).

Exercise 2.12 Expressing mutual information in terms of entropies

Show that

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2.136}$$

Exercise 2.13 Mutual information for correlated normals

(Source: (Cover and Thomas 1991, Q9.3).) Find the mutual information $I(X_1, X_2)$ where $\mathbf{X}$ has a bivariate normal distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right) \tag{2.137}$$

Evaluate $I(X_1, X_2)$ for $\rho = 1$, $\rho = 0$ and $\rho = -1$ and comment. Hint: The (differential) entropy of a d-dimensional Gaussian is

$$h(\mathbf{X}) = \frac{1}{2}\log_2\left[(2\pi e)^d \det \Sigma\right] \tag{2.138}$$

In the 1d case, this becomes

$$h(X) = \frac{1}{2}\log_2\left[2\pi e\sigma^2\right] \tag{2.139}$$

Hint: $\log(0) = \infty$.

Exercise 2.14 A measure of correlation (normalized mutual information)

(Source: (Cover and Thomas 1991, Q2.20).) Let X and Y be discrete random variables which are identically distributed (so $H(X) = H(Y)$) but not necessarily independent. Define

$$r = 1 - \frac{H(Y|X)}{H(X)} \tag{2.140}$$

a. Show $r = \frac{I(X,Y)}{H(X)}$

b. Show $0 \leq r \leq 1$

c. When is $r = 0$?

d. When is $r = 1$?

Exercise 2.15 MLE minimizes KL divergence to the empirical distribution

Let $p_{\text{emp}}(x)$ be the empirical distribution, and let $q(x|\boldsymbol{\theta})$ be some model. Show that $\text{argmin}_q \mathbb{KL}\left(p_{\text{emp}}||q\right)$ is obtained by $q(x) = q(x; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the MLE. Hint: use non-negativity of the KL divergence.

Exercise 2.16 Mean, mode, variance for the beta distribution

Suppose $\theta \sim \text{Beta}(a, b)$. Derive the mean, mode and variance.

Exercise 2.17 Expected value of the minimum

Suppose X, Y are two points sampled independently and uniformly at random from the interval $[0, 1]$. What is the expected location of the leftmost point?