2.1 Random Variables

We will need some definitions from probability theory as well as a smattering of statistics nomenclature and definitions. This section can be safely skipped by those with familiarity with those subjects. We will also set the stage for some of our notation in this section as well.

Starting in this section we will use the convention of denoting a random variable by a capital letter, e.g., X, and a *realization or sample* from that random variable using the lower case of the same letter, in this case x. In other words, x is a realization of random variable X.

2.1.1 Probability Density and Cumulative Distribution Functions

The probability density and cumulative distribution functions are key pieces of information about a random variable. Sometimes we know them, for instance when we say an input to a code has a normal distribution, and others, for example a QOI, we would like to determine. In either case, we will need to know how the two are related and the key properties of each.

For a given a real random variable $X \in \mathbb{R}$, the cumulative distribution function (CDF) is defined as

$$\begin{aligned} F_X(x) &= P(X \leq x) \qquad (2.1) \\ &= \text{The probability that the random variable } X \text{ is less than or equal to some number } x. \end{aligned}$$

Oftentimes, we will leave out the subscript on F when it is clear what random variable we are referring. One of the uses of the CDF is to find the probability that a random variable is between two numbers. From the above definition it is straightforward to see that we can find the probability that X is between a and b via subtraction

$$F_X(b) - F_X(a) = P(a < X \leq b). \qquad (2.2)$$

In this equation we note that the probability is strictly greater than a and less than or equal to b. This comes from the definition of the CDF that we used. Based on the fact that a probability must be in the closed interval $[0,1]$ we assert that

$$F_X(x) \in [0,1].$$

Also, since X is a real number we know that

$$\lim_{x\to\infty} F_X(x) = 1 \qquad \lim_{x\to-\infty} F_X(x) = 0.$$

This statement is equivalent to saying that X will take some value between negative and positive infinity. There is one more property of the CDF that we need. Namely, that the CDF is nondecreasing. One way to state this is to say

$$F_X(x+\varepsilon) \geq F_X(x) \qquad \text{for } \varepsilon > 0.$$

In other words as x increases the probability the X is less than or equal to x cannot go down. We will show some examples of CDFs later.

If X is a continuous random variable, that is, X can take any real value. We define the probability density function (PDF) as

$$f(x) = \frac{dF_X}{dx} = \text{ the density of probability at a point.} \tag{2.3}$$

Because $f(x)$ is a density, if we multiply it by a differential volume element dx we get

$$f(x)dx = \text{ probability that } X \text{ is within } dx \text{ of } x.$$

We can ”invert” the definition of the PDF to get the CDF in terms of the PDF:

$$F_X(x) = \int\limits_{-\infty}^{x} f(x')\,dx'.$$

Following this line of thinking further, we deduce that the probability X is between a and b is given by

$$P(a < X \leq b) = \int\limits_{a}^{b} f(x)\,dx = F_X(b) - F_X(a).$$

Also, using the limits of $F_X(x)$,

$$\int\limits_{-\infty}^{\infty} f(x)\,dx = 1.$$

As an example of PDF and CDF, consider the normal distribution (also known as a Gaussian distribution). This distribution has two parameters, μ and σ. As we will see later these correspond to the mean and standard deviation of the distribution. A random variable X that is normally distributed has a PDF given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.4}$$

One can show that the CDF is given by

$$F(x) = \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{2.5}$$

where $\text{erf}(x)$ is the error-function. When X is a normally distributed random variable with parameters μ and σ we denote this as $X \sim \mathscr{N}(\mu,\sigma)$.

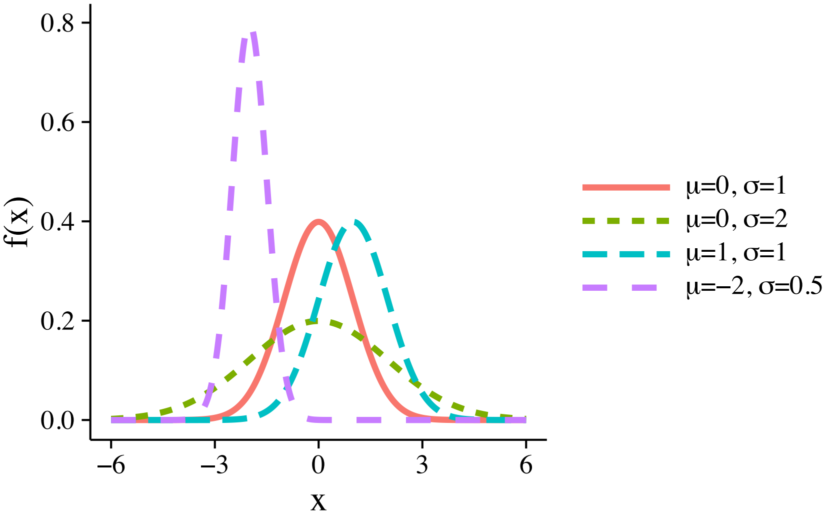

Fig. 2.1 Probability density functions for a normally distributed random variable with different values of μ and σ.

In Figs. 2.1 and 2.2 we show the PDF and CDF for a normal distribution with different values of μ and σ. Notice that the PDF is highest at $x = \mu$ and the PDF is symmetric about μ. Also, the parameter σ controls the width of the PDF with higher values making a wider PDF. From the CDF we see that $F(x)$ is equal to 0.5 at $x = \mu$, and the smaller values of σ have a steeper increase in $F(x)$. All of these features could be deduced from the definitions of the PDF and CDF.

A normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard* normal distribution and the PDF of this case is given by $\phi(x)$ and the CDF is written as $\Phi(x)$. Also, any normal random variable can be written in terms of a standard normal. If $X \sim \mathscr{N}(\mu,\sigma)$, then

$$z = \frac{x-\mu}{\sigma}, \tag{2.6}$$

will create a random variable $Z \sim \mathscr{N}(0,1)$.

2.1.2 Discrete Random Variables

For discrete random variables, that is a random variable that only take on a countable number of values, we cannot use a probability density function because it does

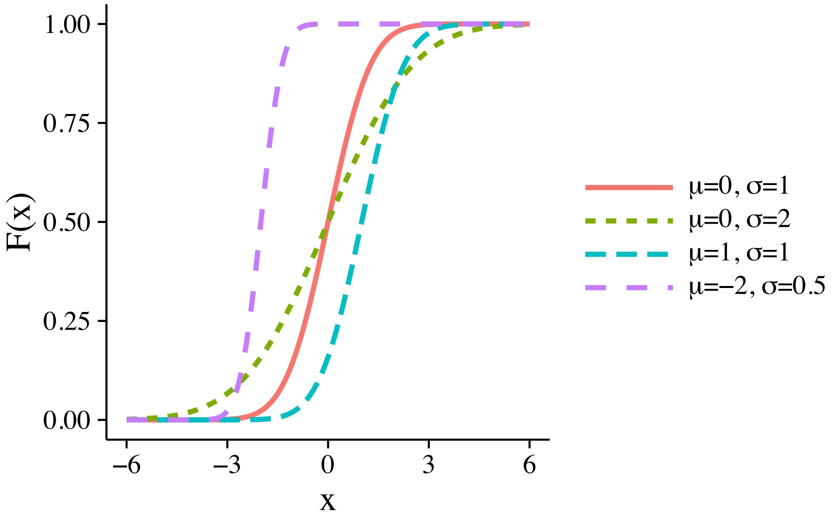

Fig. 2.2 Cumulative distribution functions for a normally distributed random variable with different values of μ and σ.

not make sense to talk about a differential volume element. Instead we define the probability mass function (PMF) for a discrete random variable as

$$f(x) = P(X = x) = \text{ the probability that } X \text{ is exactly equal to } x. \tag{2.7}$$

The notation is being somewhat abused by having both the PDF and probability mass function use f. Nevertheless, by the context it should be clear which we mean, and in practice this is a distinction without a difference if we think of the probability mass function as a sum of Dirac-delta functions. For the CDF of a discrete random variable, instead of an integral we have a sum

$$F_X(x) = \sum_{s \in S} f(s), \tag{2.8}$$

where S is the set of all possible values of X less than or equal to x.

An important example of a discrete random variable is the Bernoulli distribution, named after Jakob Bernoulli who developed it in his work *Ars Conjectandi* [Bernoulli(1713)]. This distribution is simple, but useful. It involve a random variable X that can take on two values, 0 and 1, with the probability of $x = 1$ being p. That is the PMF

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}. \tag{2.9}$$

The CDF can be easily shown to be

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases} . \tag{2.10}$$

If the random variable is a fair coin, then p is 0.5 and we can (arbitrarily) choose a flip that lands on heads as $x = 1$ and a flip that lands on tails as $x = 0$. The PMF and CDF for a Bernoulli distributed X with $p = 0.5$ is shown in Figs. 2.3 and 2.4. Notice the "stair-step" shape of the CDF because the probability that x is less than or equal to a given number "jumps" when crossing 0 and 1.

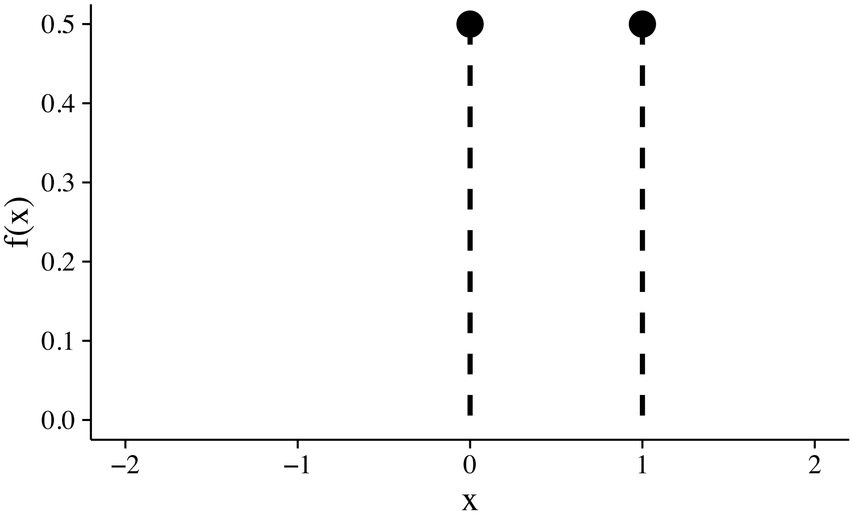

Fig. 2.3 Probability mass function for a Bernoulli distributed random variable $p = 0.5$.

2.2 Expectation Value

It is common to express properties of a random variable in terms of particular moments of its PDF or PMF called expectation values. The expectation value (or expected value) of a function $g(x)$ is denoted as $E[g(X)]$ given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \tag{2.11}$$

The expectation value is a weighted average of $g(x)$ where the weighted function is the PDF (or PMF).

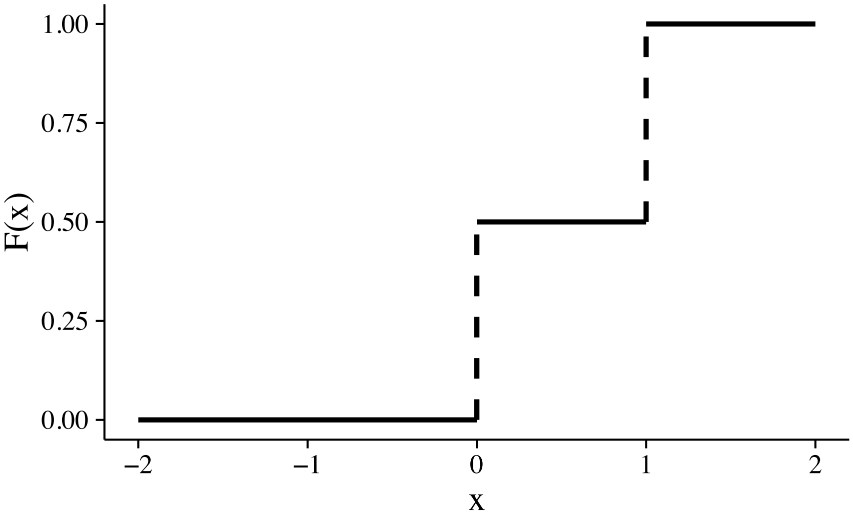

Fig. 2.4 Cumulative distribution function for a Bernoulli distributed random variable $p = 0.5$.

An important special case of the expectation value is the mean which is the expected value of x. It is often denoted as μ,

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{2.12}$$

In common parlance, the mean is the value of X one would ”expect” when drawing a random variable. In many cases this is true. For example if $X \sim \mathcal{N}(0,1)$, that is X is normally distributed with $\mu = 0$ and $\sigma = 1$. The mean of X is then

$$E[X] = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \mu. \tag{2.13}$$

The above relation can be shown by making the substitution $u = x^2$. Equation (2.13) says that μ is the mean of the distribution. It’s also true that μ is the maximum value of $f(x)$, and therefore the most likely value of X.

The mean is not always the most likely value of a random variable, in fact it may not even be a possible value of X. Consider the Bernoulli distribution, the mean of this distribution is

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = 0 \cdot (1-p) + 1 \cdot p = p. \tag{2.14}$$

Therefore, mean (or expected value of X) is p when X can only take the values of 0 or 1. The mean is still useful in this case, we just cannot interpret it as the most likely value.

An old saying about judging a random variable by its mean goes something like this: if I put my head in the oven and my feet in ice water, my mean temperature is just right. In other words, the mean does not tell us everything about the random variable.

2.2.1 Median and Mode

There are two useful properties of the distributions that are not related to the expectation value: the median and mode. The median is the point at which the CDF is equal to one-half, i.e., $F(x) = \frac{1}{2}$. This is a useful quantity because it indicates the point that splits the random variable into two equal parts: in the limit of an infinite number of realizations, half will be above the median and half will be below the median. This is not true of the mean. Also, the median is less influenced by outliers.

The mode is the point which the PDF takes its maximum value. Therefore it is the most likely value of the distribution. A distribution with a single mode is said to be unimodal.

2.2.2 Variance

The expected value of $(x-\mu)^2$ is called the variance, and often written in shorthand as σ^2. It is worth noting that the variance can be expressed in terms of the mean and $E[X^2]$ via

$$E[(X-\mu)^2] = E[X^2] - 2E[\mu X] + E[\mu^2] = E[X^2] - \mu^2.$$

In this relation, we used the fact that $E[X] = 1$, $E[\mu X] = \mu^2$, and $E[\mu^2] = \mu^2$. One can interpret the variance as the average squared difference between a random variable and its mean. The larger the value of the variance more likely values away from the mean are. The square root of the variance is called the standard deviation, σ. The standard deviation is useful because it will have the same units as X, whereas the variance has the units of X^2.

For a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma)$, the variance of X is σ^2. The fact that larger values of σ^2 correspond to values away from the mean being more likely can be seen in Fig. 2.1. In that figure those curves with larger values of σ have much wider curves. For the Bernoulli distribution, the variance can be shown to be $p(1-p)$. Therefore, the maximum value of σ^2 for the Bernoulli distribution is $0.5^2 = 0.25$ and occurs when $p = 0.5$.

2.2.3 Skewness

The mean and the variance are related to the expectation of X and X^2 respectively. The skewness, γ_1, is related to the third moment of $f(x)$, that is the expected value of X^3:

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\mathrm{Var}(X)^{3/2}}. \tag{2.15}$$

The skewness tells us something about the "tails" of the distribution, that his how the distribution goes to zero away from the mean (assuming the distribution has a single maximum (a unimodal distribution). A negative skewness tells us the distribution goes to zero more slowly to the left of the mean, whereas a positive skewness says the opposite. An example of each sign of skewness is shown in Fig. 2.5. The normal distribution has a skewness of 0 because it is symmetric about the mean.

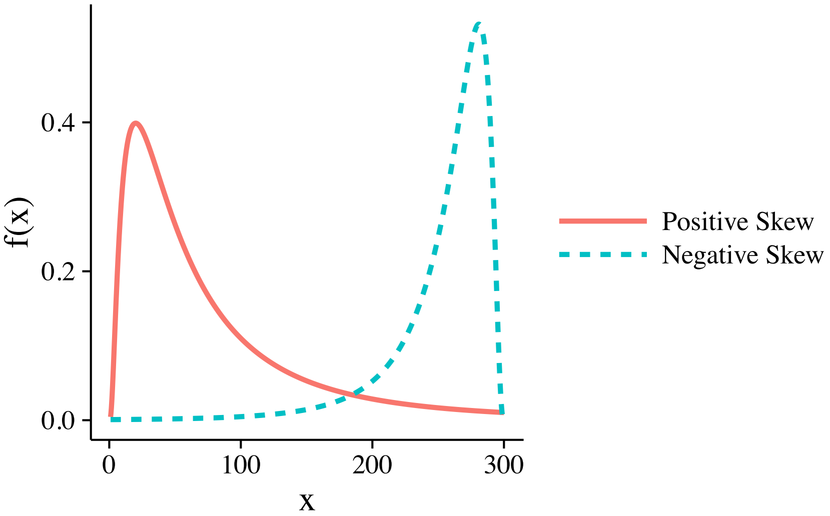

Fig. 2.5 The PDFs of two distributions demonstrating positive and negative skewness.

2.2.4 Kurtosis

Next on the list of properties of a distribution is the excess kurtosis (usually just referred to as the kurtosis) which is a measure of "tail fatness" for a distribution. The kurtosis, Kurt(X), is related to the fourth moment of a random variable's PDF and is defined as:

$$\mathrm{Kurt}(X) = \frac{E[(X-\mu)^4]}{\sigma^4} - 3. \tag{2.16}$$

The minus three is included so that a normal distribution has a kurtosis of 0. The definition of the kurtosis is such that for a unimodal distribution, the slower the PDF approaches zero as one moves away from the mode, the higher the kurtosis will be. Another way of thinking about it is that the sign of the kurtosis tells you if the distribution has heavier tails than a normal distribution (positive kurtosis) or if it has thinner tails than a normal distribution (negative kurtosis). There are also fancier names for these cases. A distribution that has negative kurtosis is said to be platykurtic from the Greek *platy* for “flat”, whereas a positive kurtosis indicates a leptokurtic distribution from the Greek word *lepto* meaning narrow. A distribution with zero kurtosis is mesokurtic.

As an example let’s look at a uniform distribution, a normal distribution, and the logistic distribution in terms of kurtosis. A uniform distribution is just like how it sounds, it has a PDF that is uniform over a finite range:

$$f_{\mathrm{uni}}(X) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}. \tag{2.17}$$

The kurtosis of a uniform distribution is $-\frac{6}{5}$, and a variance of $\frac{1}{12}(b-a)^2$. We already noted that the definition of kurtosis we are using has a normal distribution have a kurtosis of zero. The logistic distribution’s PDF is given by

$$f_{\mathrm{logistic}}(X) = \frac{1}{4s}\mathrm{sech}^2\left(\frac{x-\mu}{2s}\right), \tag{2.18}$$

where s is a parameter that acts in a similar way to the standard deviation in a normal distribution. The variance of a logistic distribution is $\frac{1}{3}s^2\pi^2$ and its kurtosis is 6/5. A uniform distribution over the range $[a,b]$ is written as $X \sim \mathscr{U}(a,b)$.

To compare these distributions we will look at each with a variance of 1. We show the three on the same plot in Figs. 2.6 and 2.7. The uniform distribution has a kurtosis of -$\frac{6}{5}$ (playtkurtic) and demonstrates this with its flat shape that approaches zero very quickly once we look far enough away from the mean. The logistic distribution is more peaked than the normal and has a positive kurtosis of $\frac{6}{5}$ (leptokurtic). In Fig. 2.6 we see the relative flatness and peakedness of these distributions. When we zoom in on the tails above $x = 3$, that is more than 3 standard deviations from the mean, we see that the leptokurtic distribution has a higher probability density than the normal distribution. This means that for the logistic distribution one is more likely to have the random variable take on “extreme values” far outside the mean.

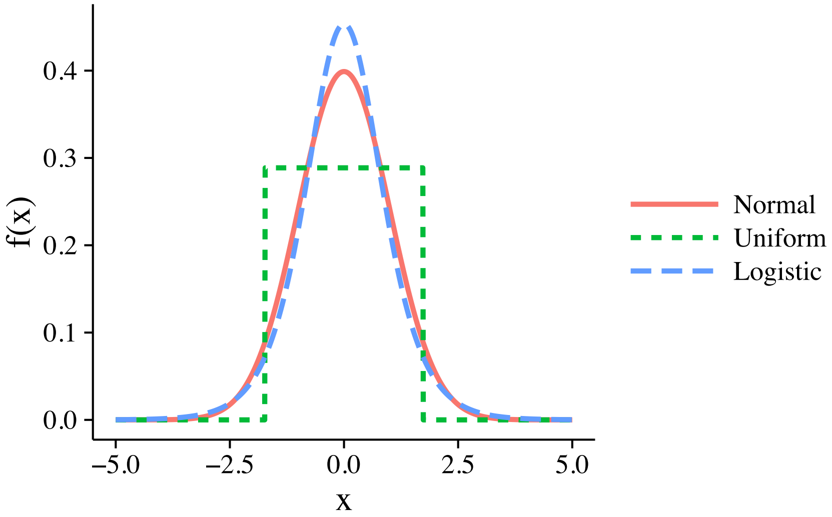

Fig. 2.6 PDFs for a uniform, normal, and logistic distribution all with mean 0 and variance 1.

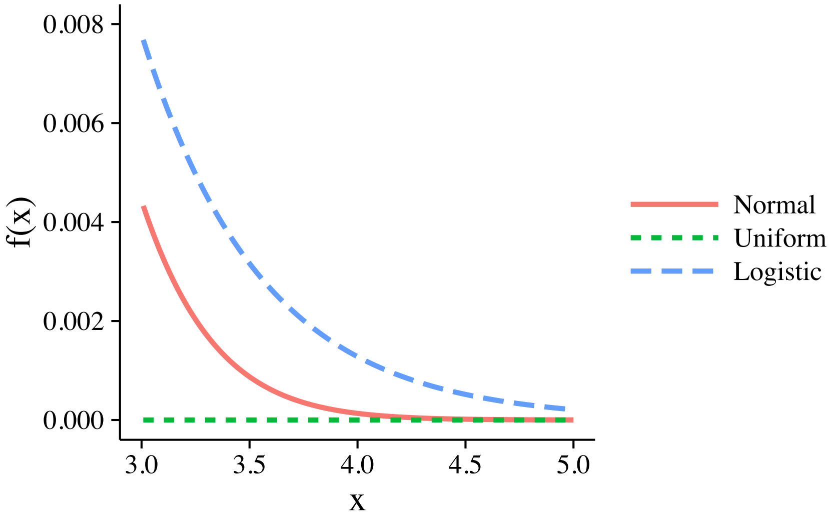

Fig. 2.7 Detail of Fig. 2.6 where we see that for the logistic distribution, one is more likely to have extreme values (greater than 3 standard deviations from the mean) than a normal distribution.

2.2.5 Estimating moments from samples

Given a number of samples, or realizations, of a random variable it is useful to estimate what the moments and other quantities of the underlying distribution are. Using this knowledge one can then approximate the probability distribution of the random variable. The moments are integrals over the probability distribution. To estimate these quantities we rely on the näive estimator:

$$E[g(x)] = \int_{-\infty}^{\infty} dx\, g(x) f(x) \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i), \tag{2.19}$$

where x_i is a sample from the PDF $f(x)$ and N is the number of samples. In other words, the expected value of $g(x)$ is approximated by the average value of $g(x_i)$. Therefore, an estimate of the mean of the PDF can be estimated via the approximation

$$\mu \approx \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \bar{x}. \tag{2.20}$$

The notation $\bar{x}$ is the used for the estimate of the mean. This estimate of the mean will have an error based on the randomness of the samples involved. One can show, via the central limit theorem, that the error is the estimate of the mean is proportional to $1/\sqrt{N}$ as $N \to \infty$.

The variance estimate is similar in that we are trying to estimate an integral. There is a slight wrinkle, however, because to estimate the variance, we use our estimate of the mean. The formula for the estimate of the variance based on a sample of random variables is written as s^2 given by:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \approx \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \approx \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \equiv s^2. \tag{2.21}$$

The factor $1/(N-1)$ comes from the fact that the we have to use the estimate of the mean, $\bar{x}$, instead of the true mean. This factor is called Bessel's correction, and comes from the fact that the quantity $(x_i - \bar{x})$ has N values but only $N-1$ independent values because the sum of $(x_i - \bar{x})$ must equal zero. Nevertheless, if N is large the correction has a small effect.

The skewness has a similar formula for an estimator, it is a combination of the sample mean, $\bar{x}$, and the sample variance, s^2, along with an additional integral estimate. The skewness estimate for a sample is written as b_1 and given by

$$b_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{(s^2)^{3/2}}. \tag{2.22}$$

The excess kurtosis for a sample is written as g_2

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{(s^2)^2} - 3. \tag{2.23}$$

There is no simple formula for computing the median of a sample. In principle, one needs to either sort the list of samples and find the middle element, if there are an odd number of samples, to take the average of the two elements adjacent to the middle of the list. There are more sophisticated algorithms that can find the smallest $N/2$ items in a list.

2.3 Multivariate Distributions

Consider a vector of random p variables: $\mathbf{X} = (X_1, X_2, \dots, X_p)^{\mathrm{t}}$. We can discuss properties of this collection of random variables in a similar way to the a single random variable. First, we define the *joint cumulative distribution function* (joint CDF) as

$$F(\mathbf{a}) = F(a_1, a_2, \dots, a_p) = P(X_1 \le a_1, X_2 \le a_2, \dots, X_p \le a_p). \tag{2.24}$$

This function is the probability that each random variable is smaller than a given number. As before, this definition allows the difference of the joint CDFs to give you the probability that each random variable is within a range:

$$F(\mathbf{b}) - F(\mathbf{a}) = P(a_1 < X_1 \le b_1, a_2 < X_2 \le b_2, \dots, a_p < X_p \le b_p).$$

As before, the derivative of the joint CDF is the *joint probability density function* (joint PDF):

$$f(\mathbf{x}) = f(x_1, x_2, \dots, x_p) = \left.\frac{\partial^p F(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_p}\right|_{\mathbf{x}}. \tag{2.25}$$

The joint CDF is then the integral of the joint PDF in a similar fashion to the single variable

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} dx_1' \int_{-\infty}^{x_2} dx_2' \dots \int_{-\infty}^{x_p} dx_p' \, f(\mathbf{x}'). \tag{2.26}$$

Using the joint PDF we can get the PDF of a single variable. For instance, $f(x_1)$ can be computed by integrating over the other $p-1$ variables:

$$f(x_1) = \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_p \, f(\mathbf{x}'). \tag{2.27}$$

That is, if we integrate over the second through p^{th} variables, we will have a function of just x_1 that is equal to it PDF. In this case we would call $f(x_1)$ the *marginal* probability density function for random variable X_1. Additionally, we can define a marginal cumulative distribution function for X_1 as

$$F(x_1) = \int_{-\infty}^{x_1} dx'_1 \int_{-\infty}^{\infty} dx'_2 \dots \int_{-\infty}^{\infty} dx'_p \, f(\mathbf{x}'). \tag{2.28}$$

Clearly, the marginal PDF and CDF could be defined for any of the p variables in the multivariate distribution.

We can generalize the idea of the marginal PDF into the joint marginal PDF of any sub-set of the p variables. Say for $l < p$ variables, the joint PDF for these l variables is

$$f(x_1, x_2, \dots, x_l) = \int_{-\infty}^{\infty} dx_{l+1} \dots \int_{-\infty}^{\infty} dx_p \, f(\mathbf{x}'). \tag{2.29}$$

These definitions then allow us to define a *conditional probability distribution function* (conditional PDF). The conditional PDF gives the distribution of a collection of random variables provided that another collection of random variables takes particular values. For an example, imagine a collection of two random variables, X and Y. We can define the probability distribution of Y provided $X = x$ as

$$f(y|X = x) = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)\, dy} = \frac{\text{Prob. density of } x \text{ and } y}{\text{Prob. density of } x \text{ for any } y}. \tag{2.30}$$

Using the definition of Eq. (2.27), we can simplify this, for $f_X(x) \neq 0$.

$$f(y|X = x) = \frac{f(x,y)}{f_X(x)},$$

where we have used the subscript X to indicate that f_X is the PDF of the random variable X. Going back to the more general case, the conditional probability of l random variables given $p - l$ other variables is

$$f(x_1, \dots, x_l | X_{l+1} = x_{l+1}, \dots, X_p = x_p) = \frac{f(\mathbf{x})}{f(x_{l+1}, \dots, x_p)} \qquad f(x_{l+1}, \dots, x_p) \neq 0, \tag{2.31}$$

where

$$f(x_{l+1}, \dots, x_p) = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_l \, f(\mathbf{x}).$$

The mean of a collection of random variables is just the vector, $\mu = (\mu_1, \dots, \mu_p)$, of the expectation values for each element in the collection:

$$\mu_i = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_p \, x_i f(\mathbf{x}). \tag{2.32}$$

The variance for a collection of random variables is more complicated than that for a single variable because we can look at how the random variables change together.

The measure of this is called the covariance and the covariance between X_i and X_j is written as σ_{ij}:

$$\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_p \, (x_i - \mu_i)(x_j - \mu_j) f(\mathbf{x}), \tag{2.33}$$

note $\sigma_{ij} = \sigma_{ji}$. The covariance of X_i with itself is the variance of X_i:

$$\sigma_{ii} = \sigma_i^2 = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_p \, (x_i - \mu_i)^2 f(\mathbf{x}). \tag{2.34}$$

The covariances form a p by p symmetric matrix with the diagonal being the variance of each random variable. The covariance matrix is typically denoted by $\Sigma(\mathbf{x})$ so that

$$\Sigma_{ij}(\mathbf{x}) = \sigma_{ij}. \tag{2.35}$$

There is a special case for a collection of random variables where the joint PDF can be factored into the product of individual PDFs as

$$f(\mathbf{x}) = \prod_{i=1}^{p} f(x_i).$$

This type of multivariate distribution is said to be independent: the value of one random variable does not depend on the value another random variable takes. An independent collection of random variables will have a covariance matrix with no non-zero off-diagonal elements. However, the opposite is not true. It is possible to have zero covariance between variables but for them to not be independent.

Example: Multivariate Normal Distribution

The multivariate normal distribution is a higher-dimension version of the normal random variable. In this case a collection of variables is jointly distributed according to a mean value for each, a covariance matrix for the relation between the variables. The probability density function for a multivariate normal PDF of k variables is given by

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \tag{2.36}$$

Here $\mathbf{x}$ is a k-dimensional vector, $\mathbf{x} = (x_1, x_2, \dots, x_k)^{\mathrm{T}}$, $\boldsymbol{\mu}$ is a vector of the expected value, or mean of each of the random variables X_i:

$$\boldsymbol{\mu} = (E[X_1], E[X_2], \dots, E[X_k])^{\mathrm{T}} = (\mu_1, \mu_2, \dots, \mu_k)^{\mathrm{T}},$$

and the covariance matrix Σ was defined in Eq. (2.35), with the determinant of the matrix written as $|\Sigma|$. The notation for a random variable $\mathbf{X}$ to be a multivariate normal with mean vector, $\boldsymbol{\mu}$, and covariance matrix Σ is $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

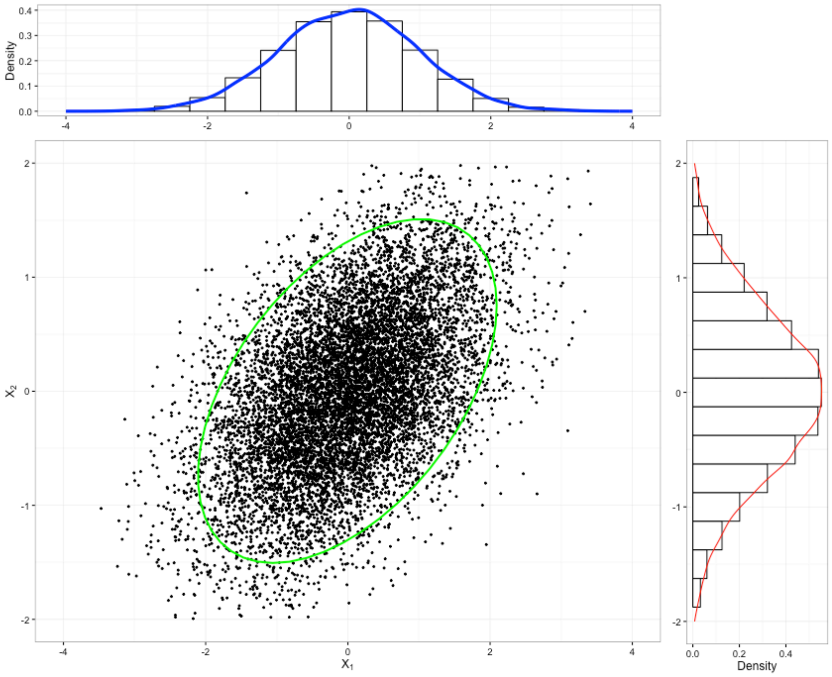

Fig. 2.8 Ten-thousand samples from a 2-D multivariate normal random variable with $\mathrm{Var}(x_1) = 1$, $\mathrm{Var}(x_2) = 0.5$ and covariance $\sigma_{12} = 0.35$. The histograms show the marginal distribution of the samples and the ellipse is the 95% probability interval for the distribution (i.e., the integral of the joint CDF over that ellipse will be 0.95).

2.4 Sampling a random variable

In general it is easy to get a random variable that is uniformly distributed between 0 and 1. In fact, almost all programming languages have a function for generating such random numbers. We would like the ability to generate a random sample of any type of random variable. This is can be done if the CDF of the distribution is known by inverting the CDF of the random variable. As mentioned above, the CDF is a function that has a range from $[0,1]$ and is a monotonic non-decreasing function. Therefore, there the CDF is invertible. With this result we can take a uniformly distributed random variable between 0 and 1, and invert the CDF to get a sample of the random variable associated with that CDF. That is,

$$x = F^{-1}(\xi), \qquad \xi \sim \mathscr{U}(0,1), \tag{2.37}$$

will give a sample x that is distributed according to the CDF $F(x)$.

An illustration of this procedure is shown in Figure 2.9 for a standard normal random variable. Here we show samples of a uniformly distributed variable between 0 and 1, and the corresponding samples from the distribution after inverting the CDF. Notice where the CDF is changing, there is a higher density of samples.

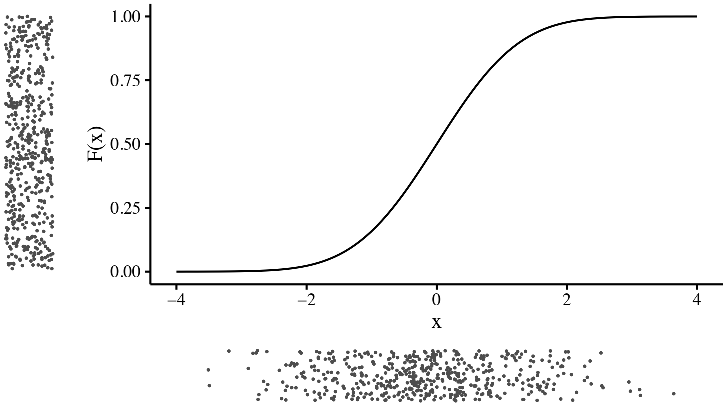

Fig. 2.9 In this figure we show a set of points on the y-axis that are randomly chosen between 0 and 1. Then on the x axis we show the corresponding sample points from inverting the CD (in this case the standard normal CDF). Notice that the uniform samples in y are non-uniformly clustered around 0, as we would expect for samples from a standard normal random variable.

Example: Sampling from an exponential PDF

An exponential random variable has PDF

$$f(x) = \lambda e^{-\lambda x}, \;\; x \geq 0,$$

where λ is a hyperparameter. The CDF is

$$F(x) = \int_0^x \lambda e^{-\lambda x'} dx' = 1 - e^{-\lambda x}$$

Choose $\xi \in [0,1]$ randomly and set

$$F(x) = 1 - e^{-\lambda x} = \xi \quad \Rightarrow \quad 1 - \xi = e^{-\lambda x}.$$

Therefore,

$$x = \frac{-\log(1-\xi)}{\lambda},$$

and x will be distributed according to $f(x)$.

Example: Normal Random Variable

In this example we will explain a clever way of inverting the CDF for a standard normal random variable. A sample from a standard normal random variable can be transformed to a general random variable through the relation,

$$x = \mu + z\sigma, \qquad Z \sim \mathcal{N}(0,1).$$

Consider a normal random variable with mean 0 and standard deviation 1. The associated PDF will be

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

The Box-Muller transform gives a way to get two samples at a time. Consider the product of two PDFs

$$f(x)\,dx f(y)\,dy = \frac{e^{\frac{-(x^2+y^2)}{2}}}{2\pi}\,dxdy.$$

If we change coordinates into polar coordinates so that

$$dxdy = r\,dr\,d\theta,$$

for $r = \sqrt{x^2+y^2}$, and $\theta = \tan^{-1}(y/x)$. Therefore, we can write

$$f(x)f(y)\,dydx = e^{\frac{-r^2}{2}}\,r\,dr\frac{d\theta}{2\pi} \quad r \in [0,\infty),\, \theta \in [0,2\pi].$$

We can separate this expression into two functions,

$$g(r) = e^{\frac{-r^2}{2}}\,r,$$

and

$$h(\theta) = \frac{1}{2\pi}.$$

These functions are both properly normalized PDF's:

$$\int_{0}^{\infty} g(r)\,dr = \int_{0}^{2\pi} d\theta\, h(\theta) = 1.$$

One can easily sample a θ:

$$\theta = 2\pi\xi_1, \qquad \xi_1 \in [0,1].$$

To sample an r from $g(r)$, we can use the result from the previous example if we define $u = r^2$ and $du = 2rdr$ to get

$$r = \sqrt{-2\log(1-\xi_2)}, \qquad \xi_2 \in [0,1].$$

As a result, drawing two random numbers, ξ_1 and ξ_2, gives two samples from the Gaussian:

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

This compares with the brute force approach of inverting the CDF for a normal random variable, with inverting two simple CDFs. The tradeoff is that one needs to generate two samples at a time.

2.4.1 Sampling a Multivariate Normal

Consider a collection of p random variables that are multivariate normal, $\mathbf{X} \sim \mathscr{N}(\boldsymbol{\mu}, \Sigma)$. To sample from this distribution we will first take the Cholesky decomposition of the covariance matrix:

$$\Sigma = \mathbf{L}\mathbf{L}^{\mathrm{T}},$$

where L is a lower-triangular matrix. The Cholesky decomposition exists for any symmetric matrix of real values that is positive definite. The covariance matrix satisfies these properties. The Cholesky decomposition requires $O(p^3)$ floating point operations to compute, and is therefore expensive when p is large.

With the Cholesky decomposition, we then generate p independent samples from a standard normal random variable:

$$\mathbf{z} = (z_1, \ldots, z_p)^{\mathrm{T}}, \qquad Z_i \sim \mathscr{N}(0,1).$$

To get a sample from $\mathbf{X}$ when then compute

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}.$$

To demonstrate how this procedure works, we look at the covariance matrix of the vector $\mathbf{Z} \sim \mathscr{N}(\mathbf{0}, \mathbf{1})$. In terms of the expected value

$$\Sigma(\mathbf{Z}) = E[\mathbf{Z}\mathbf{Z}^{\mathrm{T}}] = \mathbf{I}.$$

Now consider a vector $\mathbf{X} = \mathbf{LZ}$. The covariance matrix for this collection of random variables is

$$E[\mathbf{XX}^{\mathrm{T}}] = E[\mathbf{LZ}(\mathbf{LZ})^{\mathrm{T}}] = E[\mathbf{LZZ}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}]$$

From this we can move the **L**'s from outside the expectation operator to get

$$E[\mathbf{XX}^{\mathrm{T}}] = \mathbf{L}E[\mathbf{ZZ}^{\mathrm{T}}]\mathbf{L}^{\mathrm{T}} = \mathbf{LL}^{\mathrm{T}} = \Sigma(\mathbf{X}).$$

To shift this result to a variable with a non-zero mean, we add in the desired mean $\boldsymbol{\mu}$.

2.5 Rejection Sampling

many times it is difficult to create the CDF from the PDF or the CDF may not be known in closed corm, not be invertible except by expensive numerical solution. In this case, it can be easier to use rejection sampling. To illustrate how this works, we will take the PDF for a random variable X, where the random variable takes values only inside a given range $[a,b]$. We then draw a rectangle around the function. The base of the rectangle extends from a to b and the height of the rectangle is the maximum value of the PDF, called h here. An example of this is shown in Figure 2.10. Then we pick points at random in the box, i.e., $X \sim \mathscr{U}(a,b)$, and $Y \sim \mathscr{U}(0,h)$. If the

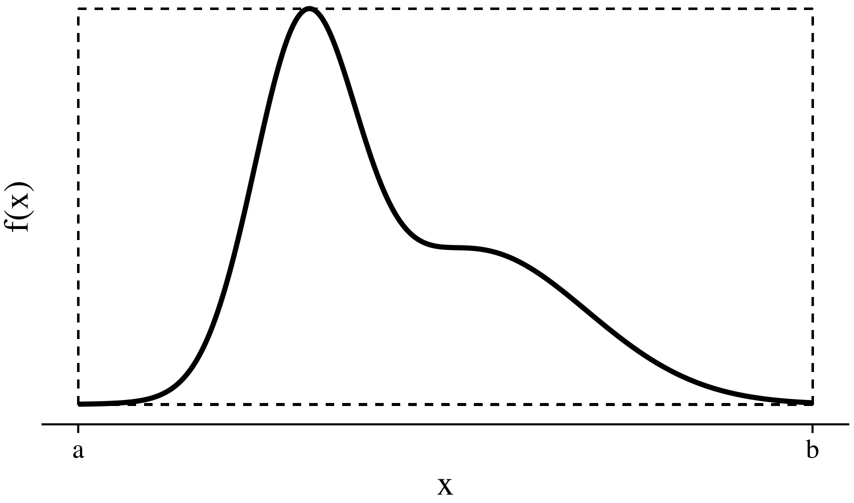

Fig. 2.10 Illustration of drawing a box around the PDF for a random variable for the purpose of rejection sampling.

point is below the PDF, then we accept it, and if not is rejected. The accepted values of X are our samples from the random variable. Figure 2.11 shows how rejection sampling proceeds as more points are tried.

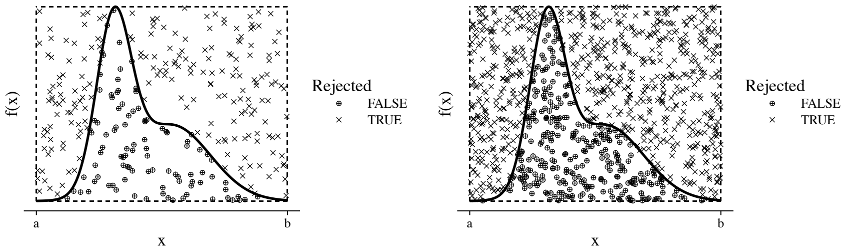

Fig. 2.11 Rejection sampling at two different number of attempted samples, 300 (left), and 1000 (right).

The rejection rate is an important measure of the effectiveness of the rejection sampling procedure. If the function is highly peaked, and goes to zero slowly, many of the sampled points can be rejected. When the rejection rate is high, it can make generating samples difficult, especially if evaluating the PDF is expensive.

One can see this in a normal random variable when the variance goes to 0. In this case the width of the PDF goes to 0. In such a case it can be more efficient to draw a different shape around the function than a rectangle. In Figure 2.12, this is demonstrated using a triangle that circumscribes the PDF. The rejection rate for this function would be much higher if we used a rectangle to bound the PDF.

2.6 Bayesian Statistics

Previously, we defined the conditional probability $f(x|Y=y)$ as the probability density that $X=x$ conditional on $Y=y$. We will often drop the"$Y=$" part from the expression. Note that we could define parameters in a distribution as a random variable. For example, the mean and variance of normal random variable could be random variables. In this sense we could write the conditional probability of X given $\mu=0$ and $\sigma^2=1$ as

$$f(x|\mu=0,\sigma^2=1)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Additionally, in Eq. (2.31) we wrote that the conditional probability was the joint probability density function divided by the marginal probability density function, viz.

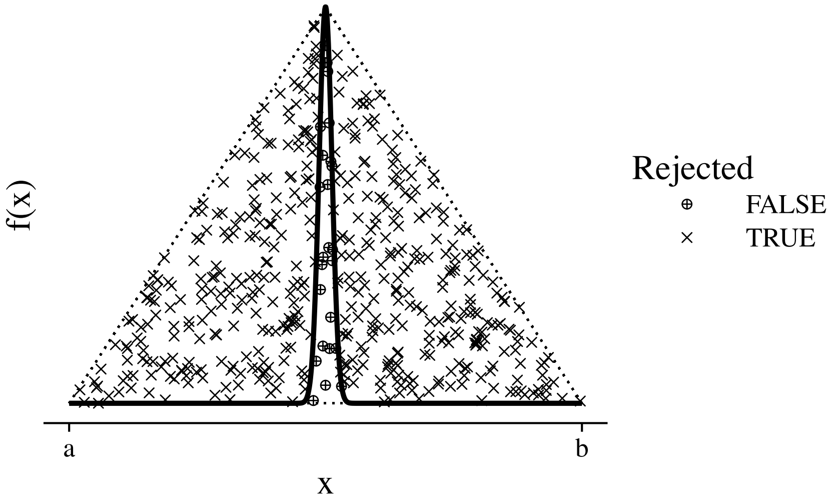

Fig. 2.12 Rejection sampling using a triangle to bound the PDF.

$$f(x_1,\ldots,x_l|X_{l+1}=x_{l+1},\ldots,X_p=x_p)=\frac{f(\mathbf{x})}{f(x_{l+1},\ldots,x_p)} \qquad f(x_{l+1},\ldots,x_p)\neq 0.$$

Therefore, for random variables X and Y the conditional probability of X given Y can be written as

$$f(x|y)f_Y(y)=f(x,y), \tag{2.38}$$

additionally, the conditional probability of Y given X can be written as

$$f(y|x)f_X(x)=f(x,y). \tag{2.39}$$

Equating these two expressions and rearranging, we can write Bayes' law (or Bayes' theorem)

$$f(x|y)=\frac{f(y|x)f_X(x)}{f_Y(y)}. \tag{2.40}$$

A more common way to write Bayes' law is to use the relation

$$f_Y(y)=\int_{-\infty}^{\infty} f(x,y)\,dx=\int_{-\infty}^{\infty} f(y|x)f_X(x)\,dx.$$

Then Bayes' law is

$$f(x|y)=\frac{f(y|x)f_X(x)}{\int_{-\infty}^{\infty} f(y|x)f_X(x)\,dx}. \tag{2.41}$$

Oftentimes, we write Bayes' law using special notation the indicates the interpretation of its implications. We define $\pi(x)$ as the *prior* probability density function for X, and $\pi(x|y)$ as the *posterior* conditional probability density function for X given $Y = y$, and $f(y|x)$ as the conditional likelihood, or just likelihood, of y given $X = x$. Using this notation we write

$$\pi(x|y) = \frac{f(y|x)\pi(x)}{\int\limits_{-\infty}^{\infty} f(y|x)\pi(x)\,dx}. \tag{2.42}$$

The interpretation of Bayes' law is that we have a prior density function for x, that we update given the observation that $Y = y$ to get $\pi(x|y)$.

Example

Assume a drug test is 99% accurate in the sense that the test will produce 99% true positive and 99% true negative. Say 0.5% of the population use the drug. An individual test spositive. What is the probability they are a user?

$$P(user|+\ test) = \frac{P(+|user)P(user)}{P(+|user)P(user) + P(+|non-user)P(non-user)}$$
$$= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} = 0.332,$$

or 33.2%.

Example

Say we want to know the fairness of a coin (i.e. is the probability of heads $\frac{1}{2}$?). If I flip the coin 10 times and get 3 heads, what is my estimate of the probability of getting heads on any toss? Using Bayes' law we write the probability of heads as p and write

$$f(p|y) = \frac{f(y|p)\pi(p)}{\int_{-\infty}^{\infty} dp f(y|p)\pi(p)}.$$

In this equation

- $f(y|p)$ = probability density of getting y given a value of p,
- $\pi(p)$ = prior distribution on p (what I believe given no data), and
- $f(p|y)$ = posterior distribution for p given data y.

For the coin example we claim to have no idea if the coin is fair, i.e. p could be anywhere between 0 and 1 with equal likelihood. We express this as

$$\pi(p)\begin{cases}1, & p\in[0,1]\\ 0, & \text{otherwise}\end{cases}$$

For the terms in Bayes' law can be filled in. The probability of getting 3 heads in 10 tosses or trials is a binomial random variable where each flip has probability p of getting heads is (see Appendix A), with probability mass function

$$f(3|p)=\binom{10}{3}p^3(1-p)^7=120p^3(1-p)^7.$$

The denominator for Bayes' theorem is

$$\int_0^1 120p^3(1-p)^7\,dp=\frac{1}{11}.$$

Putting this all together, we get the posterior

$$f(p|3)=1320p^3(1-p)^7.$$

In Figure 2.13, we show the results of this trial. We see in the posterior the maximum is at $p=0.3$, but it does not rule out the coin being fair. The posterior does rule out, however, $p=0$ or $p=1$, because those are not possible given the observation of only 3 heads.

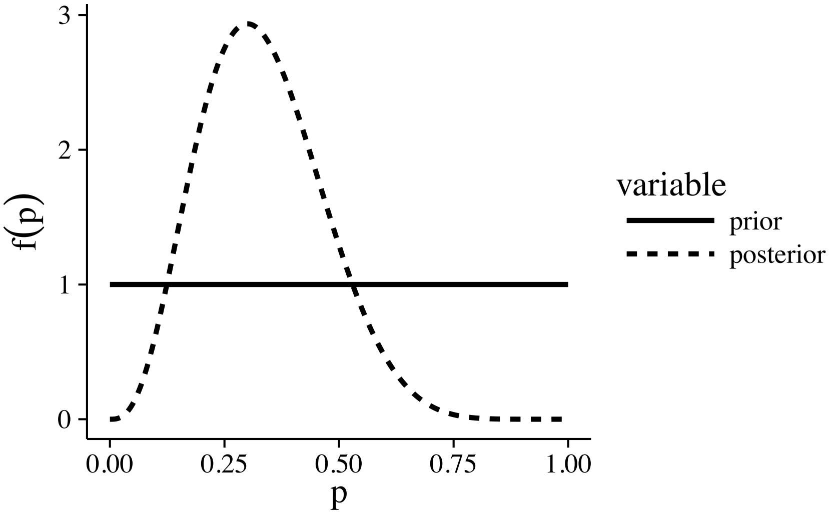

Fig. 2.13 Posterior and Prior distributions of the probability of getting 3 heads in 10 tosses for an unknown the coin tossing example.

A useful feature of Bayes' theorem is that we can update the posterior if new data comes along in the same way as before. That is, we use the current posterior as the prior in another calculation. If we make 990 more flips of the coin and 430 heads, this makes the likelihood in the numerator

$$f(430|p) = \binom{990}{430} p^{430}(1-p)^{560} = 5.127419 \times 10^{292} p^{430}(1-p)^{560},$$

and the denominator is

$$\int_0^1 1320 \binom{990}{430} p^{433}(1-p)^{567}\, dp = \frac{2016464117980615134777}{998761250084970390322850}.$$

Then, using $\pi(p) = f(p|30)$, Bayes' theorem gives

$$f(p|460) = \frac{1}{0.0020190} \binom{990}{430} p^{430}(1-p)^{560}(1320p^3(1-p)^7) \quad = \frac{1320}{0.0020190} \binom{990}{430} p^{433}(1-p)^{567}.$$

The new posterior distribution is highly peaked around $p = 0.433$, as seen in Figure 2.14, indicating that it is likely that this coin is not quite fair. The maximum of the posterior moved considerably from the result based on 10 trials. In other words, Bayes' theorem does what we would want: a large number of trials, in this case 990, has a larger impact on the posterior than a smaller number.

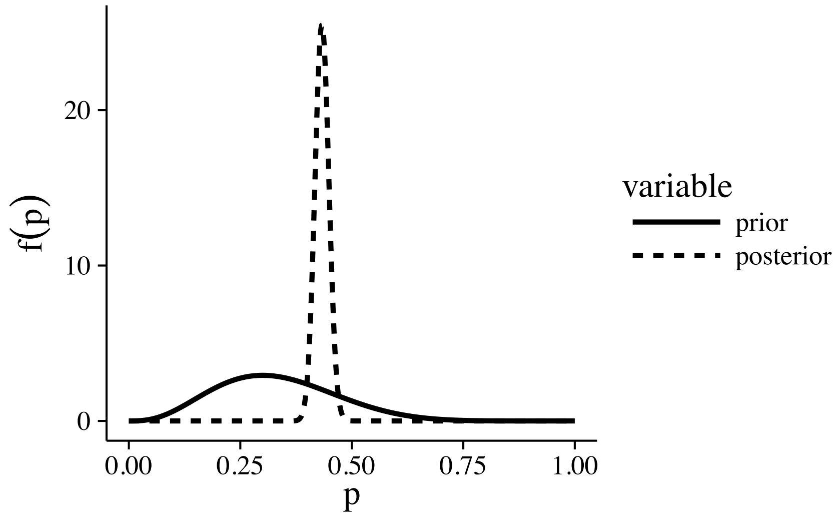

Fig. 2.14 Posterior and Prior distributions of the probability of getting heads for the coin tossing example.

In the example above we initially used what is known as an uninformed prior. That is, we put no information into the prior on p, other than saying p could be anywhere in $[0,1]$. An uninformed prior is a conservative choice when we have no other information. Had we chosen a different prior, say something peaked around $p = 0.5$, we may have gotten a different result after 10 trials.

A criticism of Bayesian calculations is that the choice of prior can matter and affect the results. Consider the hypothetical example of a shipping method of radioactive waste. If the expected value for the chance of a catastrophic accident is 10^{-3} per year of shipping, and shipping has been going on for 25 years without an accident, does that mean we can adjust the probability of a catastrophic accident? How we answer this question would depend on our choice of prior. If the prior was a delta function centered at 10^{-3}, then the distribution, and the expected failure rate would not change. However, is the distribution on the failure rate had a large variance, then the operation history would affect the posterior distribution of the failure rate.

A problem with Bayes' theorem is that it can be hard to estimate the integral in the denominator, except for some particular cases called conjugate priors. If the likelihood and the prior are chosen correctly, then the integral can be done analytically. For example, if the likelihood and the prior are both normal, then the posterior will also be normal.

Without conjugate priors, it may be difficult to estimate the integral in the denominator in Bayes' theorem. We could use quadrature approximations if we can easily evaluate the likelihood and prior. This type of approximation can too expensive if the integral is over a high-dimensional space (i.e., the x in Bayes' theorem is a vector of many variables). Later, we will discuss an approach to generate samples from a posterior distribution without needing to compute the denominator or needing a closed form for the numerator.

2.7 Exercises

1. Show that the transformation in Eq. (2.6) results in a standard normal random variable by computing the mean and variance of Z.
2. Consider the random variables $X \sim U(-1,1)$ and $Y \sim X^2$. Are these independent random variables? What is their covariance?
3. Show that a general covariance matrix must be positive definite, i.e. $\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x} > 0$ for any vector $\mathbf{x}$ that is not all zeros.
4. Use rejection sampling to sample from a Gamma random variable $X \sim \mathscr{G}(\alpha,\beta)$ where

$$f(x) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma\alpha\beta^{-\alpha}}, \qquad \alpha,\beta > 0.$$

Let $\alpha = 1$ and $\beta = 0.5$. From rejection sampling with a $N = 10^4$, compute a rejection rate for the sampling procedure. Now draw a triangle around the function and do rejection sampling. Compare the rejection rate from the triangle versus the rectangle. You may consider that the PDF is zero if $f(x) < 10^{-6}$.

5. Consider a random variable, $X > 0$, that has it's logarithm distributed by a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Such a distribution is called a log-normal distribution. Compute this distribution's a) mean, b) variance, c) median, d) mode, e) skew, and d) kurtosis.
6. (Monte Hall Problem) You are on a game show and are presented with three doors from which to choose. One of the doors contains a prize and the other two have nothing. You pick a door (say door 1), and then the host opens another door (say door 3), and asks if you want to switch to door number 2. What should you do?

    a. Using Bayes' theorem give the probability of winning if you switch.
    b. Write a simulation code to show this by randomly assigning a prize to a door, then opening either door 2 or 3 depending on which has the prize, and then either switching or not. Compute the likelihood of winning if you stick, versus the likelihood of winning if you switch.

7. Consider a variable Y distributed by a normal distribution with mean given by θ:

$$f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right).$$

    Now consider θ to be a random variable as well, and σ to be a known constant. Then say θ is normally distributed, with mean μ and variance τ^2 to give

$$\pi(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right).$$

    The parameters μ and τ are called hyperparameters. Using Bayes' theorem find $p(\theta|y)$, and show that it is a normal distribution.

8. Suppose that X is the number of people arriving at a particular tavern during a given hour. This type of arrival process is naturally described by a Poisson process:

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \qquad x \in \{0,1,2,\dots\}, \quad \theta > 0.$$

    We then say that our prior distribution of θ is a Gamma distribution

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{-\alpha}}, \qquad \theta,\alpha,\beta > 0.$$

    Therefore, we say that $\theta \sim G(\alpha,\beta)$.

    - Show using Bayes' theorem that the posterior distribution for θ given x is proportional to a Gamma distribution.

- Suppose you observe 42 people arriving in one hour, and the prior distribution has $\alpha = 5$, and $\beta = 6$. Generate samples from the posterior distribution and show graphically how the prior as changed given the observation.

9. Generate N samples from a standard normal random variable and estimate the mean, variance, skewness, and kurtosis from the samples. Use $N = 10, 10^2, \cdots, 10^4$, and discuss how the errors in the approximations behave as a function of N.
10. Consider the joint PDF

$$f(x,y) = e^{-x/y}, \qquad x \in [0,\infty) \quad y \in [0,\sqrt{2}].$$

Compute and plot the marginal PDFs for X and Y. Additionally, compute the conditional probability distributions, and make plots of $f(y|X = \mu_x)$ and $f(x|Y = \mu_y)$.