# 2

# What are random variables?

In this chapter we recapitulate some fundamental notions of probability theory, such as definitions of probability, random variables, expectations etc. We also collect a number of identities and approximations, including the delta method, which are frequently used in statistics.

## 2–1 Probability

Probability theory is a mathematical framework to describe uncertainty. For example, we might be interested in the **probability** of propositions such as "in the next toss the coin comes up with head on top" or "tomorrow it will rain". A crucial property of these propositions is that they are measurable and thus — at least hypothetically — amenable to evaluation by an **experiment**: we can throw a coin and check whether the head is on top, or we can wait until tomorrow to see whether it rains.

To formalize the above we denote by $x$ the desired outcome ("head", "rain") and rewrite the above propositions to "the outcome of the experiment equals $x$". In addition, we introduce an indicator variable $I(x)$ that takes on value 1 if the outcome $x$ is observed, and value 0 if it isn't.

With this in hand we can *define* the probability of a proposition to be the **expectation** of the indicator $I(x)$,

$$\Pr(\text{"outcome equals } x\text{"}) := \mathrm{E}(I(x)) \,. \qquad \textbf{Equation 2–1}$$

The expectation $\mathrm{E}(I(x))$ is the idealized long-running average of the outcomes in repetitions of the experiment. As $I(x)$ is either 0 or 1, it follows that a probability always lies in the interval between 0 (the outcome is never observed) and 1 (the outcome is always observed). Any number in-between implies that the outcome is random. However, it is not "random" in the colloquial sense, but rather there is a systematic pattern in the random variation of the event (namely its probability) which allows us to **predict** the frequency of outcomes in future experiments.

Note that the origin of the uncertainty plays no role whatsoever. There is no distinction (and it is indeed completely irrelevant) whether the outcome $x$ is actually random or whether it only appears to be random because of our own ignorance or inability to understand the underlying physical process. Indeed, much of the strength of probabilistic modeling stems from the fact that they allow to ignore low-level details while still capturing the essential aspects of the data generating process.

Instead of using expectation as the starting point for the definition of probability an alternative and mathematically equivalent route is to simply *postulate* probability to be a measure between 0 and 1 that is additive for disjoint propositions and then define expectation as secondary derived quantity. A third route not described here is to view probability as the unique generalization of propositional logic. This implies that that probability can be also viewed as a degree of belief. Hence, probability is defined even if you cannot repeat an experiment.

## 2–2 Random variables

We have encountered already one example of a **random variable**, namely the indicator function $I(x)$. This is a binary random variable with the set $\Omega = \{0, 1\}$ as its associated **sample space**.

More generally, a random variable $X$ is a model of a random experiment, and the set of all possible elementary outcomes comprises its sample space $\Omega$. Depending on the type of the experiment, the random variable is discrete or continuous, and the sample space contains a finite or infinite number of elements. The sample space $\Omega$ is the basis from which testable propositions are formed, such as $X = x$ or $X \in [2, 4]$. **Equation 2–1** assigns probabilities to these propositions.

For a *discrete* random variable the probability to observe outcome $x_i$

$$\Pr(X = x_i) = f(x_i)$$

is summarized by $f(x_i)$, the **probability mass function** (PMF). By construction $\sum_i f(x_i) = 1$ as the sum of the probabilities of all possible elementary events always equals one. Similarly, for a *continuous* random variable we use $f(x)$, the **probability density function** (PDF) with $\int_{-\infty}^{\infty} f(x)dx = 1$, to compute the probability

$$\Pr(x < X \leq x + dx) = f(x)dx\,.$$

The infinitesimal interval is considered instead of $x$ because for a continuous $X$ the probability of a specific point event $x$ is exactly zero (the technical reason being that it has measure zero). Note that the PDF $f(x)$ itself is *not* a probability, and can take on values larger than one.

Another way to completely characterize a random variable is to specify its **distribution**

$$F(x) = \Pr(X \leq x)\,,$$

also known as the **cumulative distribution function** (CDF). A very compact notation to state that a random variable follows a certain distribution is to write

$$X \sim F\,.$$

The CDF assumes values from from zero to one and is a monotonically increasing function. For discrete variables the CDF is given by $F(x_i) = \sum_{j \leq i} f(x_j)$, and for continuous variables by $F(x) = \int_{-\infty}^{x} f(x')dx'$.

Mathematically, the PMF/PDF and CDF both provide identical information about the uncertainty concerning $X$ and about the random variation in the outcomes (see **Figure 2–1**). In inference, however, we will find that working with distributions rather than densities is often easier and more more stable, simply because densities are harder to estimate from data than distributions. A further advantage of $F(x)$ is that it is defined in exactly the same fashion for discrete and continuous random variables.

The distribution function $F(x)$ also allows to define **quantiles**. The $\alpha$-quantile $q_\alpha$ is the $x$ value for which $F(x) = \alpha$, or

$$q_\alpha = F^{-1}(\alpha).$$

Frequently used quantiles are the **median** (the 50% quantile) and the lower and upper **quartile** ($\alpha = 1/4$ and $\alpha = 3/4$). The **interquartile range** (IQR) is the difference between the upper and lower quartile. The median and IQR are important quantities as they are measures of **location** and **scale**, respectively.

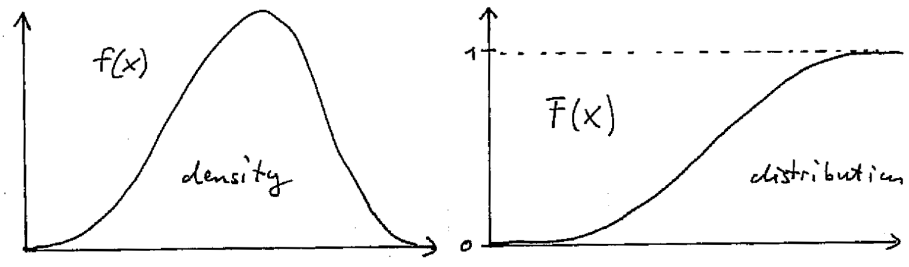


**Figure 2–1:** Comparison of density and distribution functions.

## 2–3 Expectation and moments

In **Equation 2–1** we used the expectation of the indicator random variable to define the probability of an event. Once we know the probability mass function or density we can compute the expectation of any random variable $X$ according to

$$\mathrm{E}(X) = \begin{cases} \sum_i x_i f(x_i) & \text{discrete case} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{continuous case}\,. \end{cases}$$

This formula can be generalized to the expression

$$\mathrm{E}(h(X)) = \begin{cases} \sum_i h(x_i) f(x_i) & \text{discrete case} \\ \int_{-\infty}^{\infty} h(x) f(x) dx & \text{continuous case}\,, \end{cases} \qquad \textbf{Equation 2–2}$$

where function $h(X)$ is a function of the random variable $X$. Note that these integrals not necessarily converge all the time, so we need to be aware that the expectation may not exist.

Depending on the choice of $h(x)$ we recover at a number of special cases:

- If we set $h(x) = I(x)$ we get back the probability — cf. **Equation 2–1**.
- For $h(x) = x^k$ we arrive at the $k$-th **moment** $\mathrm{E}(X^k)$.
- For $h(x) = (x - E(X))^k$ we obtain the $k$-th **centered moment**.

The **mean**

$$\mathrm{E}(X) = \mu$$

is the first moment. The **variance**

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}((X - \mu)^2)$$

is the second centered moment. If it exists, $\mathrm{Var}(X)$ is always positive and it becomes zero only if the variable is a constant. The variance can also be computed from the first and the second moment via

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2\,. \qquad \textbf{Equation 2–3}$$

The inverse of the variance is the **precision**. The square root of the variance is the **standard deviation**

$$\mathrm{SD}(X) := \sqrt{\mathrm{Var}(X)} = \sigma\,.$$

The mean is a measure of the **location** of a random variable, and the standard deviation a **measure** of the scale. This explains the practical importance of the

first two moments. However, note that in contrast to median and IQR the mean and variance need not necessarily exist.

Further quantities related to moments are $\mathrm{SD}(X)/\mathrm{E}(X)$, the **coefficient of variation** and its inverse $\mathrm{E}(X)/\mathrm{SD}(X)$, the **signal to noise ratio** (SNR) . The **variance to mean ratio** (VMR) is given by $\mathrm{Var}(X)/\mathrm{E}(X)$. Both CV and SNR are scale invariant by construction. The VMR plays an important role in characterizing families of distributions (see next chapter).

**Example 2–1:** A dice as a random variable

The space of possible outcomes for a regular dice is $\{1, 2, 3, 4, 5, 6\}$. Therefore, the random variable $X$ modeling the dice is discrete, Each of the possible outcomes has the same probability, so the probability mass function is $f(x_i) = 1/6$ for $i \in 1, \ldots, 6$. The mean of $X$ is $\mathrm{E}(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 7/2$, and the variance equals $\mathrm{Var}(X) = 35/12$.

## 2–4 Multivariate random variables

An **univariate** random variable $X$ with a one-dimensional sample space can be viewed as a special case of a **multivariate** random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$. All the univariate definitions for distributions, densities and moments are also valid in the multivariate setting. In addition, there are some notions that require two or more variables, and hence only exist in the multivariate case.

The first moment of $\boldsymbol{X}$ is the mean vector

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$$

which contains the means of the individual $X_i$. The variance of $\boldsymbol{X}$ is the $d \times d$ sized **covariance matrix**

$$\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = (\sigma_{ij}) = \mathrm{E}((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T) \,.$$

The diagonal entries $\sigma_{ii}$ of $\boldsymbol{\Sigma}$ are the component-wise variances $\mathrm{Var}(X_i) = \sigma_i^2$, whereas the off-diagonal elements $\sigma_{ij}$ are the covariances $\mathrm{Cov}(X_i, X_j)$. The covariance matrix can also be expressed as

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T \qquad \textbf{Equation 2–4}$$

which is a generalization of **Equation 2–3**. The matrix $\boldsymbol{\Sigma}$ is symmetric and in general positive definite. This means it is invertible and has only positive eigenvalues. (However, $\boldsymbol{\Sigma}$ degenerates and becomes singular if two or more variables are perfectly linear functions of each other.) The inverse $\boldsymbol{\Sigma}^{-1}$ is called the **precision matrix** or the **concentration matrix**. A multivariate measure of scale is given by the matrix

$$\mathrm{SD}(\boldsymbol{X}) := \sqrt{\mathrm{Var}(\boldsymbol{X})} = \boldsymbol{\Sigma}^{1/2} \,.$$

The matrix square root of a positive definite matrix (such as $\boldsymbol{\Sigma}$) is well defined, unique, and also positive definite. It is typically computed via spectral decomposition (see Appendix). Note that $\boldsymbol{\Sigma}^{1/2}$ is symmetric, too. The **correlation matrix**

$$\boldsymbol{P} = (\rho_{ij}) = \boldsymbol{V}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{V}^{-1/2}$$

is the standardized version of the covariance, where $\boldsymbol{V} = \text{diag}\{\sigma_1^2, \ldots, \sigma_d^2\}$ contains the variances. The correlation matrix $\boldsymbol{P}$ is invariant against scale transformations of the individual $X_i$ and has diagonal elements $\rho_{ii} = 1$ and off-diagonal entries

$$\text{Corr}(X_i, X_j) = \rho_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}}\,.$$

The correlation $\rho_{ij}$ assumes values between -1 and +1 and describes the *linear* dependence betweem $X_i$ and $X_j$. If two variables are perfectly correlated then $|\rho_{ij}| = 1$. Conversely, if they are uncorrelated then $\rho_{ij} = 0$. Written in terms of $\boldsymbol{P}$ the scale matrix $\text{SD}(\boldsymbol{X})$ decomposes into

$$\text{SD}(\boldsymbol{X}) = \boldsymbol{\Sigma}^{1/2} = \boldsymbol{V}^{1/2} \boldsymbol{P}^{1/2}\,.$$

In the multivariate setting one needs to carefully distinguish between the **joint distribution** $F_{\boldsymbol{X}}$ and the **marginal distribution** $F_{X_i}$ of the individual components. Two random variables $X_1$ and $X_2$ are called **independent** if their joint density can be written as the product of the respective marginal densities,

$$f_{\boldsymbol{X}}(x_1, x_2) \overset{X_1,\, X_2 \text{ independent}}{=} f_{X_1}(x_1) f_{X_2}(x_2)\,.$$

Thus, by comparing the joint and marginal densities is possible to determine whether two variables are independent (see later in the book in **Equation 4–7**).

**Example 2–2:** Correlation between $X$ and $X^2$

To see that correlation is *not* a measure of general dependence we consider a random variable $X$ with mean zero and compute the covariance $\text{Cov}(X, X^2)$. Using **Equation 2–4** and $\text{E}(X) = 0$ we get $\text{Cov}(X, X^2) = E(X^3) \approx 0$. Therefore, even though $X$ and $X^2$ are completely dependent, they have approximately correlation zero! This shows that it is in general not allowed to conclude from zero correlation that a pair of variables are independent.

A more general criterion than correlation for checking the independence of variables is mutual information (**Equation 4–7**).

## 2–5 Algebraic identities

Directly from the definitions of the mean and the variance we can derive the following two simple yet highly useful algebraic rules,

$$\text{E}(\boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}) = \boldsymbol{a} + \boldsymbol{b}\,\text{E}(\boldsymbol{X}) \qquad \textbf{Equation 2–5}$$

and

$$\mathrm{Var}(\boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}) = \boldsymbol{b}\,\mathrm{Var}(\boldsymbol{X})\,\boldsymbol{b}^T\,, \qquad \textbf{Equation 2–6}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors or matrices of appropriate dimension. It is *not* required that the vectors $\boldsymbol{X}$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$ have the same dimension.

**Example 2–3:** Some common identities

For a univariate random variable $X$ **Equation 2–5** and **Equation 2–6** reduce to

$$\mathrm{E}(a + bX) = a + bE(x)$$

and

$$\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)\,.$$

Considering a two-dimensional $\boldsymbol{X} = (X_1, X_2)^T$ with $\boldsymbol{a} = (a_1, a_2)^T$ and $\boldsymbol{b} = \mathrm{diag}\{b_1, b_2\}$ we can generalize the variance identity to

$$\mathrm{Cov}(a_1 + b_1 X_1, a_2 + b_2 X_2) = b_1 b_2\,\mathrm{Cov}(X_1, X_2)\,.$$

**Example 2–4:** Sum of two random variables

We consider the sum $Y = X_1 + X_2$, which corresponds to $a = 0$ and $\boldsymbol{b} = (1, 1)$. Using **Equation 2–5** and **Equation 2–6** we find for the mean

$$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$$

and the variance

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\,\mathrm{Cov}(X_1, X_2)\,.$$

Thus, while the expectation of the sum of two random variables always equals the sum of the individual expectations, this is holds true for the variance only if the variables are uncorrelated.

**Example 2–5:** Decorrelation by the Mahalanobis transformation

Consider a random vector $\boldsymbol{X}$ with covariance $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$, where $\boldsymbol{P}$ is the correlation matrix and $\boldsymbol{V}$ a diagonal matrix containing the variances. From $\boldsymbol{X}$ we now construct another random variable $\boldsymbol{Y} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X}$. What is the covariance of $\boldsymbol{Y}$?

Using **Equation 2–5** we get $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}\boldsymbol{P}^{-1/2} = \boldsymbol{P}^{-1/2}\boldsymbol{P}\boldsymbol{P}^{-1/2} = \boldsymbol{I}$. This means that the components of $\boldsymbol{Z}$ are all uncorrelated, with the variances all being equal to one. Thus, multiplication by $\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}$, a procedure referred to as Mahalanobis transformation, is a means of decorrelation. It is also known as "prewhitening" or "sphering". An alternative way to decorrelate the elements of $\boldsymbol{X}$ is the construction of principal components (see Problems).

## 2–6 Delta method

One often constructs a new random variable $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$ from another random variable $\boldsymbol{X}$ with known mean $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and variance $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. As in the previous section $\boldsymbol{X}$ and $\boldsymbol{Y}$ are allowed to be of different dimension. The **delta method** provides a means to approximate the mean and variance of $\boldsymbol{Y}$ for an arbitrary function $\boldsymbol{h}$.

A first order Taylor expansion

$$\boldsymbol{h}(\boldsymbol{X}) \approx \boldsymbol{h}(\boldsymbol{\mu}) + \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})$$

provides a linear approximation of $\boldsymbol{h}(\boldsymbol{X})$ around $\boldsymbol{\mu}$ using the Jacobi matrix $\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})$ (see the Appendix for details). If we rewrite this as $\boldsymbol{h}(\boldsymbol{X}) \approx \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$ with $\boldsymbol{a} = \boldsymbol{h}(\boldsymbol{\mu}) - \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})\boldsymbol{\mu}$ and $\boldsymbol{b} = \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})$ we can directly apply **Equation 2–5** and **Equation 2–6** to get the approximation

$$\mathrm{E}(\boldsymbol{Y}) \approx \boldsymbol{h}(\boldsymbol{\mu})$$

and

$$\mathrm{Var}(\boldsymbol{Y}) \approx \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})\, \boldsymbol{\Sigma}\, \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{\mu})^T .$$

Note that the delta method is exact when $\boldsymbol{h}$ is linear.

**Example 2–6:** Univariate delta method

It is easy to check that when $X$ is a univariate random variable and $h(x)$ a univariate function the delta method simplifies to

$$\mathrm{E}(Y) \approx h(\mu)$$

and

$$\mathrm{Var}(Y) \approx \sigma^2\, h'(\mu)^2 .$$

## 2–7 Coordinate transformations

A function $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ where $\boldsymbol{y}$ and $\boldsymbol{x}$ have the same dimension and which has a unique inverse $\boldsymbol{y} = \boldsymbol{h}^{-1}(\boldsymbol{x})$ specifies a **coordinate transformation**. In abbreviated notion we write $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ and $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})$. In this section we explore how a change of coordinates affects the distribution, probability mass and density functions of a random variable $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$.

The distribution function of $\boldsymbol{Y}$ is obtained from $F_X$ by appropriate mapping,

$$F_Y(\boldsymbol{y}) = \begin{cases} F_X\Big(\boldsymbol{x}(\boldsymbol{y})\Big) & \text{for increasing } \boldsymbol{x}(\boldsymbol{y}), \text{ and} \\ 1 - F_X\Big(\boldsymbol{x}(\boldsymbol{y})\Big) & \text{for a decreasing transformation.} \end{cases}$$

Similarly, for a discrete random variable the transformed probability mass function is given by

$$f_Y(\boldsymbol{y}_i) = f_X\Big(\boldsymbol{x}(\boldsymbol{y}_i)\Big).$$

For a continuous random variable it is a bit more complicated because we also need to ensure correct transformation of the volume element $\boldsymbol{dx}$ (see Appendix). As a result we get

$$f_Y(\boldsymbol{y}) = \left|\det\Big(\boldsymbol{J}_x(\boldsymbol{y})\Big)\right| f_X\Big(\boldsymbol{x}(\boldsymbol{y})\Big). \qquad \textbf{Equation 2–7}$$

Note the additional factor, the absolute value of the determinant of the Jacobian matrix, that is necessary for correct transformation.

**Example 2–7:** Multivariate location-scale transformation

Suppose $\boldsymbol{X}$ is a continuous random variable with known density and distribution. We apply a linear transformation $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$, where $\boldsymbol{a}$ is a location and $\boldsymbol{b}$ a scale parameter. What's the form of the distribution and density of $\boldsymbol{Y}$?

The inverse transformation is $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{b}^{-1}(\boldsymbol{y} - \boldsymbol{a})$, and the Jacobi matrix required for transformation of the density is $\boldsymbol{J}_x(\boldsymbol{y}) = \boldsymbol{b}^{-1}$. Thus, for a general location-scale transformation the distribution function of $\boldsymbol{Y}$ is

$$F_Y(\boldsymbol{y}) = F_X\Big(\boldsymbol{b}^{-1}(\boldsymbol{y} - \boldsymbol{a})\Big)$$

and the density is

$$f_Y(\boldsymbol{y}) = |\det(\boldsymbol{b})|^{-1} f_X\Big(\boldsymbol{b}^{-1}(\boldsymbol{y} - \boldsymbol{a})\Big).$$

If we set $\boldsymbol{a} = \boldsymbol{\mu}$ and $\boldsymbol{b} = \boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{\Sigma}$ is a positive definite matrix having only positive eigenvalues and thus with $\det(\boldsymbol{\Sigma}) > 0$, the above becomes

$$F_Y(\boldsymbol{y}) = F_X\Big(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})\Big)$$

and

$$f_Y(\boldsymbol{y}) = \det(\boldsymbol{\Sigma})^{-1/2} f_X\Big(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})\Big).$$

**Example 2–8:** Univariate location-scale transformation

For univariate $a$ and $b$ the transformation reduce for the distribution to

$$F_Y(y) = F_X\left(\frac{y-a}{b}\right)$$

and for the density to

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right).$$

**Example 2–9:** Transformation of quantiles

Under a change of coordinates $y = y(x)$ it is also interesting to study how the median and the lower and upper quartiles transform. More generally, we may ask how the the $\alpha$-th quantile of a random variable $X$ transforms.

By definition of the quantiles we have the identity $F_X(q_\alpha^X) = F_Y(q_\alpha^Y) = \alpha$. As $F_Y(q_\alpha^Y) = F_X(x(q_\alpha^Y))$ we get $x(q_\alpha^Y) = q_\alpha^X$, and thus find the simple relationship $q_\alpha^Y = y(q_\alpha^X)$.

## 2–8 Convolution

Using the algebraic identities from above it is straightforward to compute the expectation and variance of sums of random variables. Much more difficult is the problem of finding the corresponding distribution.

In the case of the sum of two variables $Y = X_1 + X_2$ the density of $Y$ can be obtained by reparameterizing the joint density of $X_1$ and $X_2$ in terms of $Y$ (note that $X_2 = Y - X_1$), and to subsequently marginalize out the remaining original variable. This leads to

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1 X_2}(x_1, y - x_1)\, dx_1 .$$

There is no need for a correction factor for $dy$ because there is no change of scale. For a discrete random variable the integral is replaced by summation. If $X_1$ and $X_2$ are *independent* then $f_{X_1 X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ and thus

$$f_Y(y) \overset{X_1,\, X_2 \text{ independent}}{=} \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1) dx_1 .$$

The density $f_Y(y)$ computed according to this formula is called the **convolution** of $f_{X_1}$ and $f_{X_2}$.

If we are interested in the sum of a large number of iid samples $X_1, \ldots, X_n$ with finite mean $\mu$ and finite variance $\sigma^2$ we can make use of the **central limit theorem** to determine the density of $Y = \sum_{i=1}^n X_i$ (see subsequent chapter).

## 2–9 Jensen's inequality

The relationship between the first two moments in **Equation 2–3** provides an example of **Jensen's inequality**. If $h(x)$ is a convex function (with positive second derivative) it can be shown that

$$h(\mathrm{E}(X)) \leq \mathrm{E}(h(X))\,. \qquad \textbf{Equation 2–8}$$

If $h(x)$ is concave (with negative curvature) then the inequality is reversed. For example, with $h(x) = x^2$ we get $E(X)^2 \leq E(X^2)$. In this particular case the difference is the variance of $X$.

## 2–10 Bibliographic notes

All of the material here is standard and can be found in most probability and statistics text books, for example in Davison (2003). In probability theory one can either use probability or expectation as the starting point of axiomatization. Here, we have followed the expectation route which is described in detail by Whittle (2000). For unified treatment of univariate and multivariate random variables as well as of continuous and discrete variables one needs to refer to measure theory (e.g., Pollard, 2002). The idea of viewing probability as a unique form of generalized propositional logic is due to Cox (1946) and advocated in Jaynes (2003).

## 2–11 Problems

2–1 Prove Jensen's inequality (**Equation 2–8**).

2–2 Show that computing the expectation $\mathrm{E}(h(X)) = \mathrm{E}(Y)$ using $f_Y$ yields the same result as applying **Equation 2–2**, which employs $f_X$.

2–3 A random vector $\boldsymbol{X}$ has a covariance matrix $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Using spectral decomposition it can be written as $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is an orthormal basis (i.e. $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$) and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$ a diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}$. Show that a covariance matrix is always positive definite, i.e. has only positive eigenvalues (assuming none of the variables are linear functions of each other).

2–4 Use the delta method to find approximations for the mean and variance of the product and the ratio of two random variables.

2–5 A random variable $X \sim F$ is transformed to $Y = h(X)$ using the transformation function $h(x) = 1 - F(x)$. What is the distribution of $Y$?

2–6 From a random vector $\boldsymbol{X}$ and the spectral decomposition of its covariance matrix $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ we construct a new random variable according to $\boldsymbol{Y} = \boldsymbol{U}^T\boldsymbol{X}$ (note that $\boldsymbol{U}$ is not symmetric). Show that its covariance $\text{Var}(\boldsymbol{Y})$ equals $\boldsymbol{\Lambda}$, which implies that the components of $\boldsymbol{Y}$ (also known as principal components) are uncorrelated.

2–7 Show that $\text{Var}(X)\text{Var}(Y) \geq \text{Cov}(X, Y)^2$. This relationship is known as the **covariance inequality**.

**In a nutshell**

- Probability is a number between 0 and 1. Probabilities are attached to measureable propositions, i.e. those that can (at least hypothetically) be evaluated by an experiment.
- Random variables are models of random experiments. They may be univariate or multivariate, or discrete or continuous.
- They are completely described by specification of a distribution function.
- Important characteristics are quantiles and moments. Of particular interest are the median and interquartile range, and the mean and standard deviation, which are indicators of location and scale, respectively.
- There is a variety of algebraic identities that allow calculation with expectations such as means and moments without knowing the complete distribution. In addition there are linear approximations (delta method).
- A change of variables requires particular care when transforming the density of a continuous random variable.

**Related chapters**

Distributions, Linear Model, Graphical Models