Chapter 1

Modeling uncertain outcomes

1.1 The three M's of decision making under uncertainty

Economic activities produce outcomes which are not completely known in advance. Uncertainty is inherent in most business plans, investment decisions, credit contracts or valuations for future liabilities. For correct decision making, uncertainty and risk has to be quantified by using probability models. This is the first step in the decision making process, which is composed of the three M's: *Modeling*, *Measuring* and *Managing* Risk.

Modeling consists in finding appropriate probability distributions for all uncertainty factors, which may affect the consequences of the decision. The modeling step is crucial. A wrong model selection jeopardizes the quality of the final decisions. When dealing with decision making under uncertainty, we distinguish between the ambiguity problem and the uncertainty problem:

- The *ambiguity problem* refers to the problem of correct model selection. It is mainly a problem of statistics. The model error, i.e. the error caused by the model ambiguity, has to be taken into account in the overall assessment of the final decision process.
- The *uncertainty problem* refers to the problem of making correct decisions, if the model has been selected. In solving the uncertainty problem it is assumed that the random distributions of all risk factors are known, only their outcomes are unknown.

Typically there are three different sources of information for establishing a probability model for the risk factors:

- historical data,
- theoretical considerations and
- expert opinion.

Based on historical data one may construct a probability distribution, which assigns to every observed historic value equal probability. This distribution is called *empirical distribution*. This simple way of modeling is done by many risk managers, however, it does not take into account possible trends and long-term changes.

Expert opinion is methodologically simple. Experts may be asked about giving lists of possible scenarios and possible degrees of plausibility, expressed in terms of probabilities. Opinions of several experts may be joined by averaging or weighting.

Theoretical models are based on classes of random variables or random processes developed in probability theory. Such classes are for instance stochastic differential and difference equations, vector autoregressive models, ARMA models, GARCH models and so on. Methods of statistics are used to identify these models and to define idealized scenario sets for them.

Model selection and the ambiguity problem are outside the scope of this book. It is assumed that the basic model for risk factors is already established and all further considerations about risks are based on this model.

Measuring risk comes next. There is a plethora of risk measures which have been proposed in literature. In Chapter 2, many of these risk measures are reviewed and their properties are presented. The choice of the appropriate measure of risk is crucial for the subsequent decision making step.

As an example, suppose a decision is to be made about whether to buy a car for a direct deposit of EUR 22.660 (variant A) or in three equal yearly installments of EUR 8.000 each (variant B). In a deterministic world, all what is needed is to know the market interest rates, r_1 for the first year and r_2 for the second year. Then the net present value (NPV) for variant A is 22.660, whereas the NPV for variant B is calculated according to the discounting formula

$$\text{NPV} = 8000 \cdot [1 + (1 + r_1)^{-1}(1 + (1 + r_2)^{-1})] \tag{1.1}$$

scenario number s	1	2	3	4	5	6
probability	0.18	0.17	0.16	0.16	0.16	0.17
rate value $r_2(s)$	5.0%	5.5%	6.0%	6.5%	7.0%	7.5%

Table 1.1.

scenario number s	1	2	3	4	5	6
probability	0.18	0.17	0.16	0.16	0.16	0.17
NPV of variant B	22.735	22.701	22.667	22.634	22.601	22.568

Table 1.2.

If, for instance, $r_1 = r_2 = 6\%$, then the NPV for variant B is 22.667. Thus the right decision would be to go for variant A.

However, the interest rates for the second year are not known. The interest rates may go up or down according to the general business climate. The assumption that there is more than one possible future situation leads to the notion of scenarios. A *scenario* is a possible situation of the parameters which influence the outcome of the decision. Under a discrete scenario model we understand a list of possible scenarios together with the corresponding probabilities of occurrence. For instance, a scenario model for the interest rates in the second year is given in Table 1.1.

Based on this scenario list, the NPV of variant B becomes a random variable. Its values, shown in Table 1.2, are calculated using the discounting formula (1.1).

The expected NPV of variant B is $(22.735 \cdot 0.18 + 22.701 \cdot 0.17 + 22.667 \cdot 0.16 + 22.634 \cdot 0.16 + 22.601 \cdot 0.16 + 22.568 \cdot 0.17) = 22.652$. If the expectation is the only criterion, variant B seems better than A. However, there is a risk in choosing it, since with probability larger than a half (precisely $0.18+0.17+0.16 = 0.51$), the decision maker is better off with variant A. To make a final decision one has to quantify the amount of risk in variant B in an appropriate way and to see whether this risk is acceptable.

Managing risks is the third of the three M's and the final step in the decision making process. Optimal decisions can be found on the basis of the quantified risks and the decision objectives by *stochastic optimization*. The choice of risk measures (risk functionals) determines the structure of the optimization problem (as linear, convex, combinatorial etc.). Considerations about the mathematical properties of the decision problem are

typically one important determinant in the problem of choosing the right quantification of risk. This aspect will be treated in detail in Chapters 4, 5 and 6. While Chapter 4 deals with single-stage decisions, Chapters 5 and 6 present some typical multi-stage problems in specific application areas.

1.2 Probability models and scenario distributions

Modeling random quantities as random variables requires the notion of probability spaces. A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the scenario set, $\mathcal{F}$ is a σ-algebra (σ-field) and $\mathbb{P}$ is a probability measure. We assume throughout the book that the probability space is non-atomic, even if all considered random variables take only finitely many values. Equivalently, one could say that we assume that the probability space is the standard space, i.e. the unit interval [0,1], endowed with Lebesgue measure.

For a short introduction into probability spaces, σ-algebras, measurability and the atomless property, see section A of the Appendix. Random variables Y are real valued functions defined on a probability space $Y : \Omega \mapsto \mathbb{R}$. Random variables are characterized by their distribution function

$$G(u) = \mathbb{P}\{Y \leq u\}.$$

The random variable Y is *discrete*, if its distribution function is a step function. For instance, the distribution function of the discrete random net present values of Table 1.2 is shown in Figure 1.1.

Idealized models for random variables have typically continuous distribution functions, which possess densities: g is the density of G, if

$$G(u) = \int_{-\infty}^{u} g(v)\, dv.$$

Examples of continuous distributions are the normal distribution, the log-normal distribution, the Gamma distribution, the Beta distribution etc. Figure 1.2 shows a continuous distribution (the Gamma distribution with parameters $a = 4$ and $b = 1/4$) and its density $g(u) = 42.66\ u^3 \exp(-4u)$.

Many risk models contain several random quantities, called the *risk factors*. For instance, financial models may contain returns for different asset categories or exchange rates for different currencies. Such models are

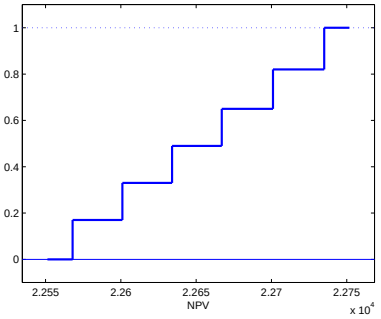

Fig. 1.1 The discrete distribution function of the data in Table 1.2.

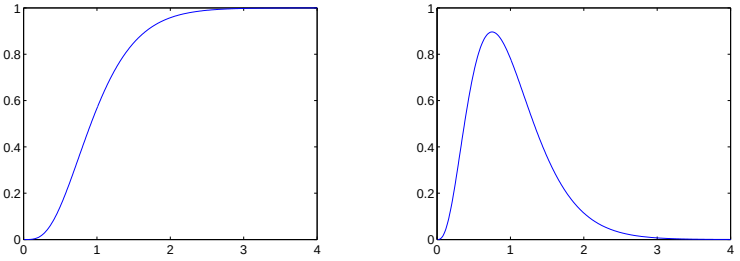

Fig. 1.2 The continuous $Gamma(4, 0.25)$ distribution function (left) and its density (right).

called *multivariate models* and the collection of several random variables is called a random vector. Random vectors are measurable mappings from Ω to $\mathbb{R}^M$. The multivariate distribution of a random vector

$$Y = (Y^{(1)}, \ldots, Y^{(M)})$$

is given by the multivariate distribution function

$$G(u_1, \ldots, u_M) = \mathbb{P}\{Y^{(1)} \leq u_1, \ldots, Y^{(M)} \leq u_M\}.$$

Risk models may also include several decision or observation periods. These models are called *multi-period models*. For such models the index t denotes time. Typically in this book the time index is an integer ranging from $t = 0$ (now) to $t = T$ (the end of the decision/observation period or the maturity time). The random variables in a multi-period model are denoted by $Y_1, \ldots, Y_T$, if the model is also multivariate by

$$Y_1^{(1)}, \ldots, Y_T^{(1)}, \ldots, Y_1^{(M)}, \ldots, Y_T^{(M)}.$$

Multivariate models are easy generalizations of univariate models and do not cause additional problems besides the problem of modeling correctly the joint distribution of all random quantities. In contrast, multi-period models are much more complicated in what concerns risk quantification. The reason is that risk is related to predictability and therefore risk functionals for multi-period models have to incorporate information and predictability. Measuring risk for multi-period models is treated in detail in Chapter 3.

1.2.1 *Distribution functions and quantile functions*

Recall that if Y is a random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, its distribution function G is

$$G(u) = \mathbb{P}\{Y \leq u\}.$$

G is continuous from the right, i.e. $G(u_n) \downarrow G(u)$, if $u_n \downarrow u$ for all u. If $u_n \uparrow u$ implies that $G(u_n) \uparrow G(u)$, then u is called a *continuity point* of G. If u is not a continuity point, we denote by

$$G(u-) = \lim_{v \uparrow u} G(v) \tag{1.2}$$

the left-sided limit of G at u.

The *quantile function* is the inverse distribution function

$$G^{-1}(p) = \inf\{u : G(u) \geq p\}, \qquad \text{for } 0 < p \leq 1. \tag{1.3}$$

Since G is continuous from the right, the infimum in (1.3) is in fact a minimum, for $0 < p < 1$. Notice that G^{-1} is continuous from the left. The following properties for distribution functions and quantile functions are well known and stated without proof.

1.1 Proposition: Properties of the quantile function.

(i) For all $0 < p < 1$

$$G(G^{-1}(p)) \geq p; \tag{1.4}$$

equality holds here, if p is in the range of G, or equivalently if $G^{-1}(p)$ is a continuity point of G.

(ii) For all $u \in \mathbb{R}$

$$G^{-1}(G(u)) \leq u; \tag{1.5}$$

equality holds here, if u is in the range of G^{-1}, or equivalently if $G(u)$ is a continuity point of G^{-1}.

(iii)

$$G^{-1}(p) \le u \quad \text{if and only if} \quad p \le G(u). \tag{1.6}$$

(iv) Suppose that Y has distribution function G and quantile function G^{-1}. The following table shows distribution functions and quantile functions of related random variables.

Y	G	G^{-1}
λY	$G(u/\lambda)$	$\lambda G^{-1}(p)$
$Y + a$	$G(u - a)$	$G^{-1}(p) + a$
$-Y$	$1 - G(-u-)$	$-G^{-1}((1-p)+)$

Here $G^{-1}(p+) = \lim_{q \downarrow p} G^{-1}(q)$.

1.2 Proposition: The quantile transform. Let the random variable Y have distribution function G. If U is a uniform [0,1] random variable, then $G^{-1}(U)$ has the same distribution as Y.

Proof. This follows from

$$\mathbb{P}\{G^{-1}(U) \le v\} = \mathbb{P}\{U \le G(v)\} = G(v)$$

which is a consequence of Proposition 1.1 (iii). □

A converse assertion does not hold: If G has jumps, then $G(Y)$ is not uniformly distributed. In general, $G(Y)$ is stochastically larger and $G(Y-)$ is stochastically smaller than a uniform [0,1] distribution. To correct for this, the generalized quantile transform must be used.

1.3 Proposition: The generalized quantile transform. Let U be a uniform [0,1] variable, which is independent of Y. Then

$$F(Y,U) := (1-U) \cdot G(Y-) + U \cdot G(Y) \tag{1.7}$$

is uniformly [0,1] distributed and $G^{-1}(F(Y,U)) = Y$ a.s.

Proof. For every $p \in (0,1)$, let y_p satisfy $G(y_p-) \le p \le G(y_p)$. y_p needs not be unique. Then

$$\mathbb{P}\{F(Y,U) \le p | Y\} = \begin{cases} 1 & \text{if } Y < y_p \\ \frac{p - G(y_p-)}{G(y_p) - G(y_p-)} & \text{if } Y = y_p \\ 0 & \text{if } Y > y_p \end{cases}$$

and consequently

$$\mathbb{P}\{F(Y,U) \le p\} = G(y_p-) + [G(y_p) - G(y_p-)] \frac{p - G(y_p-)}{G(y_p) - G(y_p-)} = p.$$

(see also [Ferguson (1967)]). To prove the second assertion, notice that conditional on $Y = u$, $F(Y,U)$ lies in the interval $[G(u-), G(u)]$ and with probability 1 in the half-open interval $(G(u-), G(u)]$. However in the latter interval, G^{-1} equals u. □

1.2.2 *Joint distributions and couplings*

For two random variables $Y^{(1)}$ and $Y^{(2)}$, the joint distribution function G_{12} is defined by

$$G_{12}(u,v) = \mathbb{P}\{Y^{(1)} \le u, Y^{(2)} \le v\}.$$

The two marginal distributions are

$$G_1(u) = \mathbb{P}\{Y^{(1)} \le u\} = G_{12}(u,\infty) \quad \text{and}$$
$$G_2(u) = \mathbb{P}\{Y^{(2)} \le v\} = G_{12}(\infty,v).$$

The knowledge of the marginal distributions is not sufficient for determining the joint distribution. The additional information needed is the copula.

1.4 Definition: Copula function. Suppose that $Y^{(1)}$ and $Y^{(2)}$ are two random variables, with distribution functions G_1 and G_2, respectively. We say that $Y^{(1)}$ and $Y^{(2)}$ have the *copula* function C, if

$$\mathbb{P}(Y^{(1)} \le u, Y^{(2)} \le v) = C(G_1(u), G_2(v)). \tag{1.8}$$

The copula function $C(s,t)$ is defined as a distribution function on $[0,1] \times [0,1]$ with uniform marginals (that is $C(1,t) = t$, $C(s,1) = s$) satisfying (1.8). For continuous marginals G_1, G_2, the copula is uniquely determined (Sklar's Theorem, see [Nelson (2006)]). Since for any sets A, B,

$$\max(\mathbb{P}(A) + \mathbb{P}(B) - 1, 0) \le \mathbb{P}(A \cap B) \le \min(\mathbb{P}(A), \mathbb{P}(B)),$$

any copula function lies between the two Fréchet bounds

$$\max(s + t - 1, 0) \le C(s,t) \le \min(s,t)$$

which implies for the distribution functions that

$$\max(G_1(u) + G_2(v) - 1, 0) \le G_{12}(u,v) \le \min(G_1(u), G_2(v)). \tag{1.9}$$

Special cases for the copula are:

(i) The *independent* case: $C(s,t) = s \cdot t$, i.e.

$$\mathbb{P}\{Y^{(1)} \le u, Y^{(2)} \le v\} = G_1(u) \cdot G_2(v). \tag{1.10}$$

(ii) The *comonotone* case (the upper Fréchet bound): $C(s,t) = \min(s,t)$, i.e.

$$\mathbb{P}\{Y^{(1)} \le u, Y^{(2)} \le v\} = \min(G_1(u), G_2(v)). \tag{1.11}$$

(iii) The *antimonotone* case (the lower Fréchet bound): $C(s,t) = \max(s+t-1,0)$, i.e.

$$\mathbb{P}\{Y^{(1)} \le u, Y^{(2)} \le v\} = \max(G_1(u) + G_2(v) - 1, 0). \qquad (1.12)$$

If the joint distribution of two random variables $Y^{(1)}$ and $Y^{(2)}$ has copula C, we say that the two random variables are *coupled* with C. Coupling is the technique to find the joint distribution, which exhibits some desired properties, when the marginals are fixed. Some coupling theorems will be presented later in this section.

Monotone resp. antimonotone couplings are the extreme cases of coupling and most different from independence: If $Y^{(1)}$ and $Y^{(2)}$ are independent, then knowing $Y^{(1)}$ does not give any information about $Y^{(2)}$. In contrast, comonotonicity (antimonotonicity) is the property that $Y^{(1)}$ gives maximal information about $Y^{(2)}$ and vice versa.

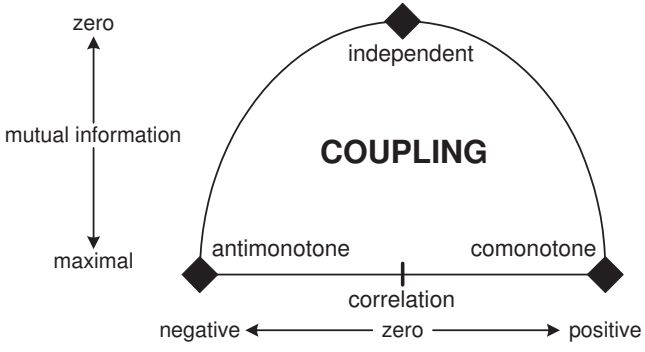

Fig. 1.3 Coupling types

1.5 Remark. If $Y^{(1)}$ and $Y^{(2)}$ are comonotone, then their common distribution is the same as the distribution of $(G_1^{-1}(U), G_2^{-1}(U))$, where U is a Uniform[0,1] variable, since

$$\begin{aligned}\mathbb{P}\{G_1^{-1}(U) \le u, G_2^{-1}(U) \le v\} &= \mathbb{P}\{U \le G_1(u), U \le G_2(v)\} \\ &= \min(G_1(u), G_2(v)).\end{aligned}$$

Similarly, if $Y^{(1)}$ and $Y^{(2)}$ are antimonotone, then their common distribution is the same as the distribution of $(G_1^{-1}(U), G_2^{-1}(1-U))$, where U is a

uniform [0,1] variable, since

$$\begin{aligned}\mathbb{P}\{G_1^{-1}(U) \le u, G_2^{-1}(1-U) \le v\} &= \mathbb{P}\{1 - G_2(u) \le U \le G_1(u)\} \\ &= \max(G_1(u) + G_2(v) - 1, 0).\end{aligned}$$

1.6 Remark. If H is monotonically increasing (resp. decreasing), then Y and $H(Y)$ are comonotone (resp. antimonotone). This follows for increasing H from

$$\mathbb{P}\{Y \le u, H(Y) \le v\} = \begin{cases} \mathbb{P}\{Y \le u\} & \text{if } H(u) \le v, \text{ i.e.} \\ & \text{if } \mathbb{P}\{Y \le u\} \le \mathbb{P}\{H(Y) \le v\} \\ \mathbb{P}\{H(Y) \le v\} & \text{if } H(u) > v, \text{ i.e.} \\ & \text{if } \mathbb{P}\{Y \le u\} \ge \mathbb{P}\{H(Y) \le v\}. \end{cases}$$

A similar assertion holds for decreasing H. However, there is no converse: Two random variables may be comonotone, but neither is a monotone function of the other.

1.7 Proposition. If $Y^{(1)}$ and $Y^{(2)}$ are comonotone, then the quantile function of $aY^{(1)} + bY^{(2)}$ for non-negative a, b is

$$aG_1^{-1}(p) + bG_2^{-1}(p).$$

If two random variables are comonotone, possess second moments and are not identically zero, their correlation is positive. Likewise, if they are antimonotone, then their correlation is negative.

Proof. Since G_1^{-1}, G_2^{-1} are left continuous and nondecreasing,

$$\mathbb{P}\{aG_1^{-1}(U) + bG_2^{-1}(U) \le aG_1^{-1}(p) + bG_2^{-1}(p)\} \ge p$$

with $U \sim$ Uniform[0,1] and for $v < aG_1^{-1}(p) + bG_2^{-1}(p)$

$$\mathbb{P}\{aG_1^{-1}(U) + bG_2^{-1}(U) \le v\} < p.$$

By the definition of the quantile function (1.3), this implies the first assertion. As to the second assertion, we invoke Hoeffding's Lemma [Hoeffding (1940)], see also [Lehmann (1966)] and Remark 2.1 in [Burgert and Rüschendorf (2006)]). If G_{12} is the joint distribution of a pair of random variables $(Y^{(1)}, Y^{(2)})$ with marginals G_1 and G_2, then the covariance $\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)})$ can be written as

$$\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) = \int\int [G_{12}(u,v) - G_1(u) \cdot G_2(v)]\, du\, dv. \tag{1.13}$$

Consequently, for comonotone $Y^{(1)}$, $Y^{(2)}$

$$\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) = \int\int [\max(G_1(u), G_2(v)) - G_1(u) \cdot G_2(v)]\, du\, dv$$

which is positive, if at least one of the variables is nondegenerated. The statement for antimonotone random variables follows by replacing $Y^{(2)}$ with $-Y^{(2)}$. □

The following example shows that although the correlation coefficient of comonotone pairs must be positive, it may be arbitrarily small.

Example. Let for $a > 1$

$$(Y^{(1)}, Y^{(2)}) = \begin{cases} (1, a) & \text{with probability } \frac{1}{2a^2} \\ (1, 0) & \text{with probability } \frac{1}{2} - \frac{1}{2a^2} \\ (-1, 0) & \text{with probability } \frac{1}{2} - \frac{1}{2a^2} \\ (-1, -a) & \text{with probability } \frac{1}{2a^2} \end{cases}$$

Both variables have expectation 0 and variance 1. They are comonotone and their correlation is $1/a$, which can be made arbitrary small by choosing a large.

The *coupling* problem consists in constructing a joint distribution with some given property, when the marginal distributions are given, that is to find the right copula. Here are some examples of coupling problems.

1.8 Proposition. For given marginals G_1 and G_2, the covariance is maximized, if the two random variables are comonotone coupled, i.e. if (1.11) holds. The covariance is minimized, if they are antimonotone coupled, i.e. if (1.12) holds.

Proof. We use Hoeffding's Lemma (1.13) again. By (1.9),

$$\begin{aligned} \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) &= \int\int [G_{12}(u, v) - G_1(u) \cdot G_2(v)]\, du\, dv \\ &\leq \int\int [\min(G_1(u), G_2(v)) - G_1(u) \cdot G_2(v)]\, du\, dv \end{aligned}$$

and

$$\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) \geq \int\int [\max(G_1(u) + G_2(v) - 1, 0) - G_1(u) \cdot G_2(v)]\, du\, dv,$$

which implies the assertion. □

The generalization for multivariate Y's is more complicated.

1.9 Proposition. Let $Y^{(1)} = (Y_1^{(1)}, \ldots, Y_K^{(1)})$ and $Y^{(2)} = (Y_1^{(2)}, \ldots, Y_K^{(2)})$ be two random vectors. Consider the problem to maximize

$$\sum_{k=1}^{K} \mathbb{E}[Y_k^{(1)} Y_k^{(2)}] \tag{1.14}$$

among all joint distributions, such that vector $Y^{(1)}$ has distribution G_1 and vector $Y^{(2)}$ has distribution G_2. If one may find a convex function f with conjugate function f^* such that

$$\sum_{k=1}^{K} Y_k^{(1)} \cdot Y_k^{(2)} = f(Y^{(1)}) + f^*(Y^{(2)}) \qquad \text{a.s.}$$

and if $Y^{(1)}$ resp. $Y^{(2)}$ have the correct marginals, then this pair solves the maximization problem (1.14).

Proof. For a proof see [Rüschendorf and Rachev (1990)]. □

There is unfortunately no constructive way to find the convex pair (f, f^*).

1.2.3 *Utility functions and order relations*

The concept of utility functions goes back to ([v. Neumann and Morgenstern (1944)]). The idea behind utility is that a numerical value v of an economic quantity is not conceived by all decision makers in the same way. Decision makers judge according to their individual preference and express this by an individual "utility transformation" $U(v)$.

For stochastic outcomes, the use of the expected utility was proposed and thoroughly studied by Arrow and (independently) by Pratt ([Arrow (1971); Pratt (1964)]). For a given utility function U, any two outcome variables, say $Y^{(1)}$ and $Y^{(2)}$, can be compared for their expected utility: $Y^{(2)}$ is preferred to $Y^{(1)}$, if $\mathbb{E}[U(Y^{(2)})] \geq \mathbb{E}[U(Y^{(1)})]$. Thus utility functions introduce a total ordering for outcome variables.

Utility functions are however artifacts and it is difficult to choose them in the proper way. It is much easier to agree on a certain set of utility functions and introduce a partial ordering (a preference relation) instead of the total ordering: If $\mathcal{U}$ is a class of functions on $\mathbb{R}$, then $Y^{(2)}$ is preferred to $Y^{(1)}$ w.r.t. $\mathcal{U}$, if

$$\mathbb{E}[U(Y^{(2)})] \geq \mathbb{E}[U(Y^{(1)})] \qquad \text{for all } U \in \mathcal{U}.$$

The following sets of utility functions are typically considered:

- $\mathcal{U}_{FSD}$, the set of all nondecreasing functions,

- $\mathcal{U}_{CCD}$, the set of all concave functions,
- $\mathcal{U}_{SSD}$, the set of all concave, nondecreasing functions,
- $\mathcal{U}_{CXD}$, the set of all convex functions.

Notice the following inclusions

$$\mathcal{U}_{SSD} \subseteq \mathcal{U}_{FSD}, \qquad \mathcal{U}_{SSD} \subseteq \mathcal{U}_{CCD}. \tag{1.15}$$

To these classes of utility functions, there correspond order relations for outcome distributions. We formulate these order relations in two ways: in terms of random variables and in terms of distribution functions.

1.10 Definition: Dominance.

(i) *Stochastic dominance of the first order.*

The r.v. $Y^{(1)}$ is dominated by $Y^{(2)}$ in first order sense,	The distribution G_1 is dominated by G_2 in first order sense,
$Y^{(1)} \prec_{FSD} Y^{(2)}$,	$G_1 \prec_{FSD} G_2$,
if $\mathbb{E}[U(Y^{(1)})] \le \mathbb{E}[U(Y^{(2)})]$	if $\int U(v)\, dG_1(v) \le \int U(v)\, dG_2(v)$

for all nondecreasing integrable U, i.e. for all $U \in \mathcal{U}_{FSD}$ for which the integrals exist.

(ii) *Stochastic dominance of the second order.*

The r.v. $Y^{(1)}$ is dominated by $Y^{(2)}$ in second order sense,	The distribution G_1 is dominated by G_2 in first order sense,
$Y^{(1)} \prec_{SSD} Y^{(2)}$,	$G_1 \prec_{SSD} G_2$,
if $\mathbb{E}[U(Y^{(1)})] \le \mathbb{E}[U(Y^{(2)})]$	if $\int U(v)\, dG_1(v) \le \int U(v)\, dG_2(v)$

for all nondecreasing concave integrable U, i.e. for all $U \in \mathcal{U}_{SSD}$ for which the integrals exist.

(iii) *Concave dominance.*

The r.v. $Y^{(1)}$ is dominated by $Y^{(2)}$ in the concave order sense,	The distribution G_1 is dominated by G_2 in the concave order sense,
$Y^{(1)} \prec_{CCD} Y^{(2)}$,	$G_1 \prec_{CCD} G_2$,
if $\mathbb{E}[U(Y^{(1)})] \le \mathbb{E}[U(Y^{(2)})]$	if $\int U(v)\, dG_1(v) \le \int U(v)\, dG_2(v)$

for all concave integrable U, i.e. for all $U \in \mathcal{U}_{CCD}$ for which the integrals exist.

(iv) *Convex dominance.*

The r.v. $Y^{(1)}$ is dominated by $Y^{(2)}$ in the convex order sense,	The distribution G_1 is dominated by G_2 in the convex order sense,
$Y^{(1)} \prec_{CXD} Y^{(2)}$,	$G_1 \prec_{CXD} G_2$,
if $\mathbb{E}[U(Y^{(1)})] \leq \mathbb{E}[U(Y^{(2)})]$	if $\int U(v)\, dG_1(v) \leq \int U(v)\, dG_2(v)$

for all convex integrable U, i.e. for all $U \in \mathcal{U}_{CXD}$ for which the integrals exist.

Obviously, $G_1 \prec_{FSD} G_2$ implies that $G_1 \prec_{SSD} G_2$. Similarly, $G_1 \prec_{CCD} G_2$ implies that $G_1 \prec_{SSD} G_2$. The relation $\prec_{CXD}$ is also known under the names *Bishop-de Leeuw ordering* or *Lorenz dominance*, the reason for the latter name is given in Proposition 1.16. More details about order relations can be found in [Müller and Stoyan (2002)].

Examples. Let G_1, G_2 be the distribution functions of normally distributed random variables with means μ_1 resp. μ_2 and standard deviations σ_1 resp. σ_2. If $\mu_1 \leq \mu_2$ and $\sigma_1 = \sigma_2$, then $G_1 \prec_{FSD} G_2$, but $G_1 \not\prec_{CCD} G_2$. In contrast, if $\mu_1 = \mu_2$ and $\sigma_1 > \sigma_2$, then $G_1 \prec_{CCD} G_2$, but $G_1 \not\prec_{FSD} G_2$. In both cases $G_1 \prec_{SSD} G_2$.

1.11 Remark. If ψ and ϕ are nondecreasing and concave, then so is their composition $\psi \circ \phi$. Thus $Y^{(1)} \prec_{SSD} Y^{(2)}$ implies $\phi(Y^{(1)}) \prec_{SSD} \phi(Y^{(2)})$ for all nondecreasing concave ϕ. $Y^{(1)} \prec_{SSD} Y^{(2)}$ implies also $Y^{(1)} + a \prec_{SSD} Y^{(2)} + b$ for $a \leq b$.

Before we give alternative characterizations of these order relations, let us introduce the integrated distribution function and the integrated quantile function.

1.12 Definition. The IDF and the IQF. Let G be a distribution function and let G^{-1} be its inverse, the quantile function. Suppose that the first moment of G exists.

- The *integrated distribution function* (IDF) $\mathcal{G}$ is

$$\mathcal{G}(u) = \int_{-\infty}^{u} G(v)\, dv. \tag{1.16}$$

- The *integrated quantile function* (IQF) $\mathcal{G}^{[-1]}$ is

$$\mathcal{G}^{[-1]}(p) = \int_0^p G^{-1}(q)\, dq. \tag{1.17}$$

Notice that $\mathcal{G}^{[-1]}$ is *not* the inverse of $\mathcal{G}$, we express this by setting $[-1]$ in brackets. Notice also that

$$\mathcal{G}^{[-1]}(1) = \int u\, dG(u) = \mathbb{E}\{G\}.$$

The IDF and the IQF are related by *Young's inequality*

$$\mathcal{G}(u) + \mathcal{G}^{[-1]}(q) \leq uq \quad \text{for } u \in \mathbb{R}, q \in (0,1) \tag{1.18}$$

where equality holds if and only if $G(u) = q$. The validity of this inequality can be easily seen from Figure 1.4: The sum of the two areas 1 and 2 is less than or equal to uq.

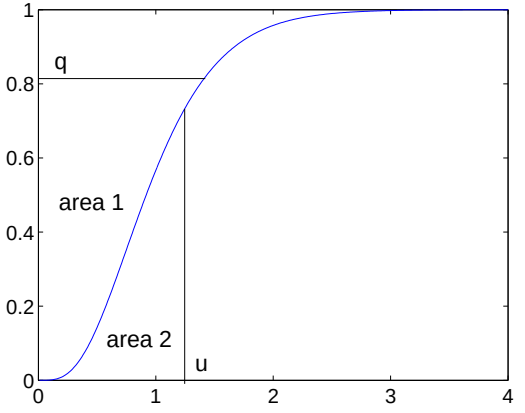

Fig. 1.4 Illustration of Young's inequality

As a consequence of (1.18) one readily sees that $\mathcal{G}$ and $\mathcal{G}^{[-1]}$ are mutually dual in the following sense:

$$\mathcal{G}(u) = \sup\{uq - \mathcal{G}^{[-1]}(q) : q \in (0,1)\} \tag{1.19}$$

and

$$\mathcal{G}^{[-1]}(q) = \sup\{uq - \mathcal{G}(u) : u \in \mathbb{R}\}. \tag{1.20}$$

We are now ready to state several alternative formulations of the order relations.

1.13 Theorem.

(i) $G_1 \prec_{FSD} G_2$ is equivalent to $G_1(u) \geq G_2(u)$ for all $u \in \mathbb{R}$.

(ii) $G_1 \prec_{FSD} G_2$ is also equivalent to $G_1^{-1}(p) \leq G_2^{-1}(p)$ for all $p \in (0,1)$.

(iii) $G_1 \prec_{CCD} G_2$ is equivalent to $G_1 \prec_{SSD} G_2$ and at the same time $\mathbb{E}\{G_1\} = \mathbb{E}\{G_2\}$.

(iv) $G_1 \prec_{CCD} G_2$ is also equivalent to the existence of a Markov kernel K with the properties

$$G_1(u) = \int K(u|v)\, dG_2(v) \quad \text{and} \quad \int u\, dK(u|v) = v \quad \text{for all } v.$$

(v) $G_1 \prec_{SSD} G_2$ is equivalent to

$$\mathcal{G}_1(u) \geq \mathcal{G}_2(u)$$

for all $u \in \mathbb{R}$.

(vi) $G_1 \prec_{SSD} G_2$ is also equivalent to

$$\mathcal{G}_1^{[-1]}(p) \leq \mathcal{G}_2^{[-1]}(p)$$

for all $p \in (0,1)$.

Proof. (i) can be easily seen since $1 - G(u) = \int \mathbb{1}_{\{v>u\}}\, dG(v)$ and $v \mapsto \mathbb{1}_{\{v>u\}}$ is nondecreasing. (ii) is a direct consequence of (i). (iii) Let $G_1 \prec_{CCD} G_2$. Then, since the identity $u \mapsto u$ is concave and convex at the same time, $\mathbb{E}\{G_1\} = \mathbb{E}\{G_2\}$. Since monotonic, concave functions are concave, $G_1 \prec_{SSD} G_2$ follows. To prove the opposite relation assume that $\mathbb{E}\{G_1\} = \mathbb{E}\{G_2\}$ and $G_1 \prec_{SSD} G_2$. We have to show that $\mathbb{E}[U(Y^{(1)})] \leq \mathbb{E}[U(Y^{(2)})]$ for every (integrable) concave function U. Recall the notion of the supergradient of a concave function U at a point v: $\partial U(v) = \{a : U(v) + a(w - v) \geq U(w) \quad \text{for all } w\}$. Call the function U supergradient-bounded from below, if $c := \inf\{a \in \partial U(v) : v \in \mathbb{R}\}$ is not equal to $-\infty$. If $c \geq 0$, then U is nondecreasing. If U is supergradient-bounded from below, then $U(w) - cw$ is nondecreasing and concave. Thus according to the assumption, $\mathbb{E}[U(Y^{(1)})] - c\mathbb{E}[Y^{(1)}] \leq \mathbb{E}[U(Y^{(2)})] - c\mathbb{E}[Y^{(2)}]$, whence $\mathbb{E}[U(Y^{(1)})] \leq \mathbb{E}[U(Y^{(2)})]$. If U is not supergradient-bounded from below, it may be approximated by a sequence of supergradient-bounded concave functions $U^{(n)}$ such that $U^{(n)} \downarrow U$ and $\mathbb{E}(U^{(n)}(Y^{(i)})) \downarrow \mathbb{E}(U(Y^{(i)}))$ for $i = 1, 2$. Since $\mathbb{E}[U^{(n)}(Y^{(1)})] \leq \mathbb{E}[U^{(n)}(Y^{(2)})]$ for each n, the same relation holds also for the limit U.

(iv) This result has many fathers: Hardy, Littlewood, Polya, Blackwell,

Stein, Sherman, Cartier, Fell, Meyer, for a proof see [Strassen (1965)].
(v) Since $\mathcal{G}(u) = \int_{-\infty}^{u} G(v)\, dv = \int_{-\infty}^{\infty} [u-v]^+ \, dG(v)$, one sees that $\mathcal{G}_1(u) \geq \mathcal{G}_2(u)$ is equivalent to $\int_{-\infty}^{\infty} \psi(v)\, dG_1(v) \leq \int_{-\infty}^{\infty} \psi(v)\, dG_2(v)$ for all functions of the form $\psi(v) = \sum_k (-\alpha_k)[u_k - v]^+ + \beta_k$, with $\alpha_k \geq 0$. These functions are dense in the set of all nondecreasing, concave functions.
(vi) The equations (1.19) and (1.20) show that the relations $\mathcal{G}_1(u) \geq \mathcal{G}_2(u)$, for all u and $\mathcal{G}_1^{[-1]}(q) \leq \mathcal{G}^{[-1]}(q)$ for all q, are equivalent (see also [Ogryczak and Ruszczyński (1999)]). □

The next example illustrates the just defined order relations for two random variables.

Example. Consider the family of Gamma distributions $Gamma(a, b, m)$ with densities

$$g_{a,b,m}(u) = \begin{cases} \frac{1}{b^a\, \Gamma(a)} (u-m)^{a-1} \exp(-(u-m)/b) & \text{for } u > m \\ 0 & \text{for } u \leq m \end{cases}$$

Let $Y^{(1)} \sim Gamma(4, 0.25, 0)$, $Y^{(2)} \sim Gamma(4, 0.25, 0.5)$, $Y^{(3)} \sim Gamma(4, 0.15, 0.5)$, $Y^{(4)} \sim Gamma(4, 0.1, 0.5)$. Then

$$\begin{gathered} Y^{(1)} \prec_{FSD} Y^{(2)}, \\ Y^{(1)} \prec_{SSD} Y^{(3)}, \text{ but } Y^{(1)} \not\prec_{FSD} Y^{(3)}, \\ Y^{(1)} \not\prec_{FSD} Y^{(4)} \text{ and } Y^{(1)} \not\prec_{SSD} Y^{(4)}. \end{gathered}$$

These relations are illustrated in Figures 1.5 – 1.7.

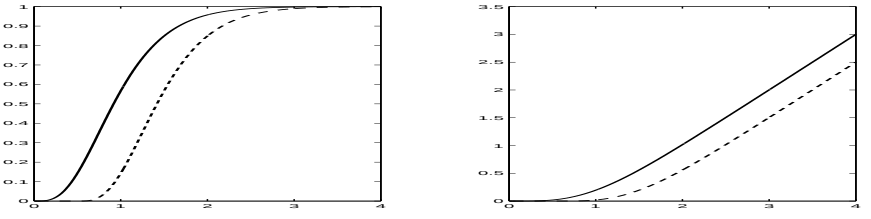

Fig. 1.5 Left: the distribution functions G_1 of $Y^{(1)}$ (solid) and G_2 of $Y^{(2)}$ (dashed) Right:the integrated distribution functions $\mathcal{G}_1$ of $Y^{(1)}$ (solid) and $\mathcal{G}_2$ of $Y^{(2)}$ (dashed); The relation $Y^{(1)} \prec_{FSD} Y^{(2)}$ holds.

The dominance relations can be concretized by some coupling. Recall that coupling is the construction of specific joint distributions (i.e. specific copulas), when the marginals are fixed.

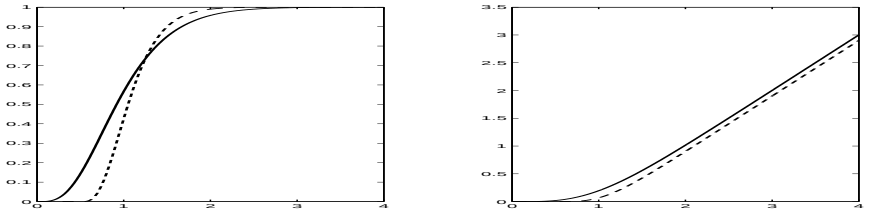

Fig. 1.6 Left: the distribution functions G_1 of $Y^{(1)}$ (solid) and G_3 of $Y^{(3)}$ (dashed) Right:the integrated distribution functions $\mathcal{G}_1$ of $Y^{(1)}$ (solid) and $\mathcal{G}_3$ of $Y^{(3)}$ (dashed); $Y^{(1)} \prec_{SSD} Y^{(3)}$ holds, but $Y^{(1)} \prec_{FSD} Y^{(3)}$ does not hold.

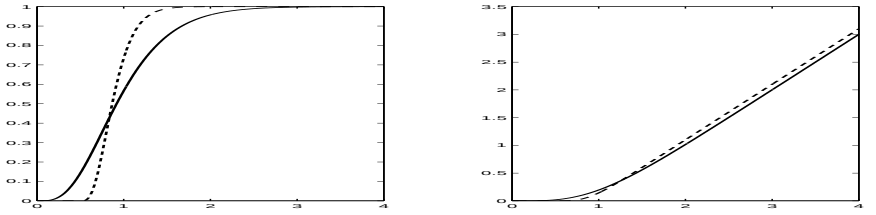

Fig. 1.7 Left: the distribution functions G_1 of $Y^{(1)}$ (solid) and G_4 of $Y^{(4)}$ (dashed) Right:the integrated distribution functions $\mathcal{G}_1$ of $Y^{(1)}$ (solid) and $\mathcal{G}_4$ of $Y^{(4)}$ (dashed); Neither $Y^{(1)} \prec_{FSD} Y^{(4)}$ nor $Y^{(1)} \prec_{SSD} Y^{(4)}$ holds.

1.14 Proposition.

(i) The FSD-coupling: If $G_1 \prec_{FSD} G_2$, then one may construct a pair $Y^{(1)}, Y^{(2)}$ of random variables with marginals G_1, G_2, such that

$$Y^{(1)} \leq Y^{(2)} \qquad a.s.$$

(ii) The CCD-coupling: If $G_1 \prec_{CCD} G_2$, then one may construct a pair $Y^{(1)}, Y^{(2)}$ of random variables with marginals G_1, G_2, such that

$$Y^{(2)} = \mathbb{E}(Y^{(1)}|Y^{(2)}) \qquad a.s.$$

(iii) The SSD-coupling. If $G_1 \prec_{SSD} G_2$, then one may construct a pair $Y^{(1)}, Y^{(2)}$ of random variables $Y^{(1)}, Y^{(2)}$ with marginals G_1, G_2, such that

$$Y^{(2)} \geq \mathbb{E}(Y^{(1)}|Y^{(2)}) \qquad a.s.$$

Alternatively, one may also construct random variables $Y^{(1)}, Y^{(2)}$ with marginals G_1, G_2 and a third random variable Y such that

$$Y^{(1)} \leq \mathbb{E}(Y^{(2)}|Y) \qquad a.s.$$

Proof. The construction for (i) is to set $(Y^{(1)}, Y^{(2)}) = (G_1^{-1}(U), G_2^{-1}(U))$, for a U uniform in [0,1]. For the constructions (ii) and (iii) see [Strassen (1965)]. □

1.15 Definition. The *Lorenz curve* pertaining to a distribution G with finite expectation is defined as

$$L_G(p) = \mathcal{G}^{[-1]}(p)/\mathcal{G}^{[-1]}(1) = \mathcal{G}^{[-1]}(p)/\mathbb{E}\{G\},$$

where $\mathcal{G}^{[-1]}$ is the IQF (see (1.17)).

For a distribution with only non-negative values, the value $L_G(p)$ of the Lorenz curve indicates the percentage of the total outcome, which is attributed to the p percent of worst scenarios. In this case, L_G is a non-negative, monotonic, convex function, with $L_G(0) = 0$, $L_G(1) = 1$, see Figure 1.8.

If the distribution may take negative values, the Lorenz curve is no longer monotonic and non-negative, but still is convex, see Figure 1.9.

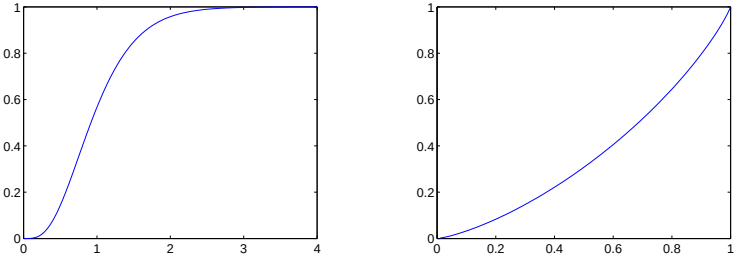

Fig. 1.8 Left: the distribution function $Gamma(4, 0.25, 0)$ Right: the pertaining Lorenz curve

Lorenz curves are related to the convex ordering by the following relationship.

1.16 Proposition. $G_1 \prec_{CCD} G_2$ is equivalent to $L_{G_1}(p) \leq L_{G_1}(p)$ for all $p \in (0,1)$ and at the same time $\mathbb{E}\{G_1\} = \mathbb{E}\{G_2\}$.

Proof. Combine (iii) and (vi) of Theorem 1.13. □

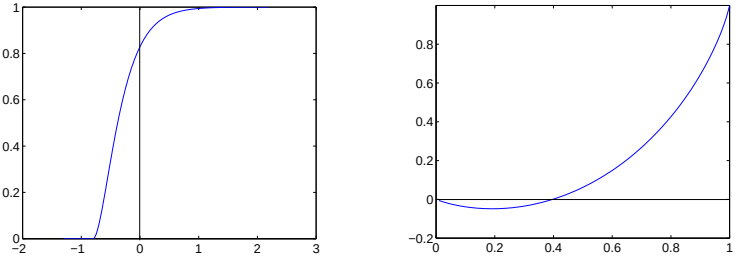

Fig. 1.9 Left: the distribution function $Gamma(2, 0.25, -0.8)$ Right: the pertaining Lorenz curve

1.2.4 *Compounding*

Loosely spoken, a compound distribution is a mixture of other distributions. For an exact definition, let $K(\cdot|u)$ be a Markov kernel, i.e. a family of distribution functions indexed with a real parameter u, and let $G(u)$ be a further distribution function. The *compound distribution* function $K \circ G$ is defined as

$$(K \circ G)(v) = \int K(v|u)\, dG(u). \tag{1.21}$$

On the level of random variables, we say that V is a compound random variable, if there is a stochastic process $(Y^{(u)})$ such that each $Y^{(u)}$ has distribution $K(\cdot|u)$ and a switching variable X with distribution G. If X is constructed on a new, independent probability space, then

$$V = Y^{(X)}$$

has distribution $K \circ G$.

The simplest compounding switches only between two variables. Let $Y^{(1)}$ and $Y^{(2)}$ be random variables and let X be the switching variable, which is independent from $Y^{(1)}, Y^{(2)}$ and which takes the value $X = 1$ with probability λ and $X = 2$ with probability $1 - \lambda$. Then the compound variable is

$$\mathcal{C}(Y^{(1)}, Y^{(2)}, \lambda) = \begin{cases} Y^{(1)} & \text{if } X = 1 \\ Y^{(2)} & \text{if } X = 2 \end{cases} \tag{1.22}$$

The distribution function of $\mathcal{C}(Y^{(1)}, Y^{(2)}, \lambda)$ is

$$\lambda G_1(u) + (1 - \lambda) G_2(u).$$

Notice the difference between compounding and comonotone coupling: If $Y^{(1)}$ and $Y^{(2)}$ are compounded with probability $1/2$, then the distribution function of the compound variable $\mathcal{C}(Y^{(1)}, Y^{(2)}, 1/2)$ is the average of the distribution functions G_1 and G_2

$$\frac{1}{2}G_1(u) + \frac{1}{2}G_2(u).$$

In contrast, if one takes the comonotone average of $Y^{(1)}$ and $Y^{(2)}$, i.e. finds comonotone versions $Y^{(1)\prime}$ and $Y^{(2)\prime}$ and takes $\frac{1}{2}Y^{(1)\prime} + \frac{1}{2}Y^{(2)\prime}$, then this random variable's quantile function is the average of the quantile functions G_1^{-1} and G_2^{-1}

$$\frac{1}{2}G_1^{-1}(p) + \frac{1}{2}G_1^{-1}(p),$$

according to Proposition 1.7. Thus compounding and monotone coupling are in a sense dual operations.

1.3 Standard statistical parameters

Since probability distributions are complex objects, there is a need of describing them with few relevant parameters. Such parameters like the mean, the variance or higher moments have been used in statistics for more than 100 years. Some of these parameters have also been used in the earlier days of risk management. However, specific requirements for appropriately measuring risk have led to new statistical parameters (new risk functionals) as they will be introduced in Chapter 2.

Some traditional statistical parameters still have their important role in risk measuring. They will be reviewed below and their properties will be discussed. We distinguish between location parameters and dispersion parameters. Notice that statistical parameters, are - by definition - only dependent on the distribution of the random variable, that is they are version independent in the sense of Definition 2.1 in Chapter 2.

1.3.1 *Location parameters*

A location parameter $\mathcal{L}$ follows all translations of the distribution, i.e. for all constant c

$$\mathcal{L}(Y + c) = \mathcal{L}(Y) + c \qquad \mathcal{L}\{G(\cdot - c)\} = \mathcal{L}\{G\} + c.$$

This property is also called *translation equivariance* (compare Definition 2.2 of Chapter 2).

Examples of frequently used location parameters are

- the *expectation* $\mathbb{E}(Y) = \int_{-\infty}^{\infty} u\, dG(u)$;
- the *median* $\mathbb{M}\mathrm{ed}(Y) = [G^{-1}(0.5) + G^{-1}(0.5+)]/2$;
- the α-quantile $G^{-1}(\alpha)$, also called the value-at-risk $\mathbb{V}@\mathrm{R}_\alpha$ of level α;
- Linear combinations of quantiles (order statistics, L-statistics)
$$\int G^{-1}(u)\; dH(u)$$
for some monotonic function H. These functionals are also called distortion functionals (see Section 2.4.2).

1.3.2 *Dispersion parameters*

A dispersion parameter $\mathcal{D}$ is unaffected by translations of the underlying distribution, i.e. for all c

$$\mathcal{D}(Y + c) = \mathcal{D}(Y) \qquad\qquad \mathcal{D}\{G(\cdot - c)\} = \mathcal{D}\{G\}.$$

This property is also called *translation invariance* (compare Definition 2.2 of Chapter 2).

Some, but not all dispersion parameters are *scale-equivariant*, i.e. they have the property

$$\mathcal{D}(\lambda Y) = |\lambda|\mathcal{D}(Y).$$

Examples are

- The *variance*.
$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y - \mathbb{E}Y)^2.$$
In terms of the distribution function, the variance can be written as
$$\mathbb{V}\mathrm{ar}\{G\} = \int (u - \int v\, dG(v))^2\; dG(u) = \frac{1}{2}\int\int (u-v)^2\; dG(u)\, dG(v),$$
which has the interpretation as
$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}(Y - Y')^2,$$
where Y' is an independent copy of Y (i.e. has the same distribution as Y but is independent of it).

- The *standard deviation*.

$$\mathbb{S}\text{td}(Y) = \sqrt{\mathbb{E}(Y - \mathbb{E}Y)^2}$$

  is the square root of the variance.
- The *mean absolute deviation*.

$$\mathbb{M}\text{ad}(Y) = \mathbb{E}|Y - \mathbb{E}Y|.$$

- The *Gini measure*.

$$\mathbb{G}\text{ini}(Y) = \frac{1}{2}\mathbb{E}|Y - Y'|$$

  where Y' is an independent copy of Y. Notice that

$$\begin{aligned}
\mathbb{G}\text{ini}(Y) &= \frac{1}{2}\mathbb{E}[Y + Y' - 2\min(Y, Y')] \\
&= \mathbb{E}Y - 2\int_{-\infty}^{\infty}\int_{-\infty}^{x} y \; dG(y) \; dG(x) \\
&= \mathbb{E}\{G\} - 2\int_{-\infty}^{\infty}\int_{0}^{G(x)} G^{-1}(p) \; dp \, dG(x) \\
&= \mathbb{E}\{G\} - 2\int_{0}^{1}\int_{0}^{q} G^{-1}(p) \; dp \, dq \\
&= \mathbb{E}\{G\} - 2\int_{0}^{1} \mathcal{G}^{[-1]}(q) \; dq \\
&= \mathbb{E}\{G\}\Big[1 - 2\int_{0}^{1} L_G(q) \; dq\Big] \qquad (1.23)
\end{aligned}$$

  where $\mathcal{G}^{[-1]}$ and L_G are the integrated quantile function and the Lorenz curve, respectively. In words, the Gini measure is the expectation minus twice the area between the identity and the Lorenz curve in the unit square.
- The *lower semi variance*

$$\mathbb{V}\text{ar}^{-}(Y) = \mathbb{E}([Y - \mathbb{E}Y]^{-})^2$$

  where $[a]^{-} = -\min(a, 0)$.
- The *upper semi variance*

$$\mathbb{V}\text{ar}^{+}(Y) = \mathbb{E}([Y - \mathbb{E}Y]^{+})^2$$

  where $[a]^{+} = \max(a, 0)$. Notice that

$$\mathbb{V}\text{ar}(Y) = \mathbb{V}\text{ar}^{+}(Y) + \mathbb{V}\text{ar}^{-}(Y).$$

- The *lower semi standard deviation*

$$\mathbb{S}\mathrm{td}^-(Y) = \sqrt{\mathbb{E}([Y - \mathbb{E}Y]^-)^2}$$

  is the square root of the lower semi variance.
- The *upper semi standard deviation*

$$\mathbb{S}\mathrm{td}^+(Y) = \mathbb{S}\mathrm{td}^-(-Y) = \sqrt{\mathbb{E}([Y - \mathbb{E}Y]^+)^2}$$

  is the square root of the upper semi variance.

The variance and the semi variances are special cases of higher central moments, resp. higher partial moments. The k-th *central absolute moment* of the distribution of Y is defined as

$$\mathbb{E}|Y - \mathbb{E}Y|^k.$$

The k-th *lower partial moment* is

$$\mathbb{E}([Y - \mathbb{E}Y]^-)^k$$

and the corresponding *upper partial moment* is

$$\mathbb{E}([Y - \mathbb{E}Y]^+)^k.$$

Partial moments have been introduced in risk management in [Bawa (1975)].

1.3.3 *Correlation parameters*

The joint behavior of two or more random variables is determined by the marginal distributions and the copula function. The covariance and correlation are statistical parameters, which give some, but not a complete information about the dependence of two random variables with second moments.

The *covariance* between the two random variables $Y^{(1)}$ and $Y^{(2)}$ is defined as

$$\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) = \mathbb{E}[(Y^{(1)} - \mathbb{E}Y^{(1)})(Y^{(2)} - \mathbb{E}Y^{(2)})],$$

while their *correlation* is

$$\mathbb{C}\mathrm{orr}(Y^{(1)}, Y^{(2)}) = \frac{\mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)})}{\mathbb{V}\mathrm{ar}(Y^{(1)})\mathbb{V}\mathrm{ar}(Y^{(2)})}.$$

For a random vector $(Y^{(1)}, \ldots, Y^{(M)})$ of length M, the *covariance matrix* is defined as

$$\begin{pmatrix} \mathbb{V}\mathrm{ar}(Y^{(1)}) & \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) & \cdots & \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(M)}) \\ \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(2)}) & \mathbb{V}\mathrm{ar}(Y^{(2)}) & \cdots & \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(M)}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(Y^{(1)}, Y^{(M)}) & \mathbb{C}\mathrm{ov}(Y^{(2)}, Y^{(M)}) & \cdots & \mathbb{V}\mathrm{ar}(Y^{(M)}) \end{pmatrix}.$$

The covariance matrix C of a random vector $(Y^{(1)}, \ldots, Y^{(M)})$ is symmetric and positive semidefinite. Such a matrix always has the representation as

$$C = V \cdot \Lambda \cdot V^T,$$

where V is an orthonormal matrix (the matrix of eigenvectors) and Λ is a diagonal matrix with non-negative entries (the eigenvalues).

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M) = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & \lambda_M \end{pmatrix}.$$

Let $V = (v_1, \ldots, v_M)$, i.e. let v_m be the columns of V. The v_m's are mutually orthogonal and have unit length. The matrix C satisfies

$$C = \sum_{m=1}^{M} \lambda_m \, v_m v_m^{\top}.$$

The transformed variables

$$Z^{(i)} = \sum_{m=1}^{M} v_{m,i} Y^{(m)}$$

are called the *principal components* of $(Y^{(1)}, \ldots, Y^{(M)})$. The covariance matrix of the principal components is Λ, i.e. the principal components are uncorrelated.

The m-th eigenvalue λ_m equals the variance of the m-th principal component $Z^{(m)}$. The proportion of the total variance which is explained by the m-th principal component is $\lambda_m / \sum_i \lambda_i$.

An example for a covariance matrix, an eigenvalue decomposition and the pertaining principal components can be found in Appendix C.