

Data Science for Insurance

Risk, Risk Factors and Loss Distributions

Roberta Pappadà

a.a. 25-26

rpappada@units.it

Table of Contents

- 1** Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2** Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3** Final comments

Table of Contents

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3 Final comments

1 Risk

- Definitions
- Dealing with uncertainty
- Extremes matter!

2 Loss distributions

- General Definitions
- Conditional vs Unconditional loss distribution
- Aggregate loss in Insurance

3 Final comments

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)
- from Wikipedia:

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)
- from Wikipedia:
 - Cambridge Dictionary: “the possibility of something bad happening”

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)
- from Wikipedia:
 - Cambridge Dictionary: “the possibility of something bad happening”
 - Oxford English Dictionary (3rd Ed): “(Exposure to) the possibility of loss, injury, or other adverse or unwelcome circumstance; a chance or situation involving such a possibility”

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)
- from Wikipedia:
 - Cambridge Dictionary: “the possibility of something bad happening”
 - Oxford English Dictionary (3rd Ed): “(Exposure to) the possibility of loss, injury, or other adverse or unwelcome circumstance; a chance or situation involving such a possibility”
 - ISO Guide 73:2009 - Vocabulary: “effect of uncertainty on objectives”

Several definitions have been adopted

- “possibility of loss or injury” (first known use of risk, 1655)
- from Wikipedia:
 - Cambridge Dictionary: “the possibility of something bad happening”
 - Oxford English Dictionary (3rd Ed): “(Exposure to) the possibility of loss, injury, or other adverse or unwelcome circumstance; a chance or situation involving such a possibility”
 - ISO Guide 73:2009 - Vocabulary: “effect of uncertainty on objectives”
- “an insurance hazard from a specified cause or source” (Merriam Webster)

- Market Risk

- Market Risk

risk associated with fluctuations in value of traded assets

- Credit Risk

- Market Risk

risk associated with fluctuations in value of traded assets

- Credit Risk

risk associated with uncertainty that debtors do not meet contractual obligations, e.g., interest on a bond is not paid

- Operational Risk

- Market Risk

risk associated with fluctuations in value of traded assets

- Credit Risk

risk associated with uncertainty that debtors do not meet contractual obligations, e.g., interest on a bond is not paid

- Operational Risk

risk associated with possibility of human error, IT failure, dishonesty, natural disaster etc.

This is a non-exhaustive list!

Another important category is, e.g, environmental risk

1 Risk

- Definitions
- Dealing with uncertainty
- Extremes matter!

2 Loss distributions

- General Definitions
- Conditional vs Unconditional loss distribution
- Aggregate loss in Insurance

3 Final comments

Risk strongly relates to uncertainty in many different situations

- Much of finance is concerned with financial risk and distributional properties of returns
- Asset pricing and risk evaluation techniques rely heavily on tools borrowed from probability theory
- Risk-management systems are based on relatively crude statistical models for the loss distribution
- Joint extreme values in several risk factors represent one of the major concerns in financial and insurance context

- Future returns from an investment cannot be known exactly, thus risk concerns the event that the investment could earn less than the expected return or result in a loss

- Future returns from an investment cannot be known exactly, thus risk concerns the event that the investment could earn less than the expected return or result in a loss
- An insurance policy sold by a company may or may not be triggered by the underlying event covered

- Future returns from an investment cannot be known exactly, thus risk concerns the event that the investment could earn less than the expected return or result in a loss
- An insurance policy sold by a company may or may not be triggered by the underlying event covered
- *“Banks are potentially exposed to climate-related financial risks regardless of their size, complexity or business model”*

Basel Committee on Banking Supervision, Nov 2021

Let X be a one-period risky position (or simply risk). Then, X is a *random variable*, that is, a function on the probability space (Ω, \mathcal{F}, P) . The range of possible values of X is R_X .

The probability that by the end of the period under consideration, the value of the risk X will be less than or equal to a given number x is the cumulative distribution function (d.f.)

$$F_X(x) = P(X \leq x)$$

Time can be introduced, leading to the notion of *stochastic process*, $\{X_t\}$, defined as a time series model for a single risk factor

Several risky positions form a random vector $\mathbf{X} = (X_1, \dots, X_d)$, defined on (Ω, \mathcal{F}, P) and taking values on \mathbb{R}^d ($d \geq 2$).

A multivariate stochastic model is represented by means of the d -dimensional (cumulative) d.f. that describes the behavior of the random vector \mathbf{X} :

$$F_{\mathbf{X}}(x_1, \dots, x_d) := F(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d)$$

The choice of $F_{\mathbf{X}}$ is a crucial task in many applied contexts where risk estimates are of interest

In fact, the d.f. for a random vector contains the description of

- 1 the marginal behavior (the probabilistic knowledge of the single components)

The choice of $F_{\mathbf{X}}$ is a crucial task in many applied contexts where risk estimates are of interest

In fact, the d.f. for a random vector contains the description of

- 1 the marginal behavior (the probabilistic knowledge of the single components)
A number of different distributions are used to model marginal components (Gaussian, Weibull, Pareto, Exponential, etc.)
- 2 the association or dependence structure between the X_i 's

The choice of $F_{\mathbf{X}}$ is a crucial task in many applied contexts where risk estimates are of interest

In fact, the d.f. for a random vector contains the description of

- 1 the marginal behavior (the probabilistic knowledge of the single components)

A number of different distributions are used to model marginal components (Gaussian, Weibull, Pareto, Exponential, etc.)

- 2 the association or dependence structure between the X_i 's
Interdependencies among risks have been modelled with simplified assumptions (e.g., independence) and/or numerical quantities (e.g., correlation coefficients) presenting well-known fallacies

Many real-world situations can be described by multivariate stochastic models.

- Portfolio Management: X_i 's can represent (daily) returns on several assets of a portfolio of investments
- Credit risk: X_i 's can represent lifetimes (time-to-default) of financial institution exposed to some shock
- Insurance: X_i 's represent potential losses in different lines of business for an insurance company
- Environmental science: natural hazards are described in terms of two or more random quantities related to the same event (e.g., storm intensity-duration, flood peak-volume, etc.)

Aggregated risks

Let

$$Y = \sum_{i=1}^d w_i X_i$$

denote the change in value of a portfolio consisting of d underlying investments over a given holding period (*profit and loss* (P/L)):

- X_i denotes the change in value of the i -th investment (e.g., rates of returns of stock prices and indices)
- weights w_1, \dots, w_d are such that $\sum_{i=1}^d w_i = 1$

So generally, 'the higher value of X_1, X_2, \dots , the better'

Measuring the risk of a portfolio essentially consists of determining its d.f. $F_Y(y) = P(Y \leq y)$, or functionals describing this function

A common way to measure the risk of Y is to look, for instance, at the lower γ -quantile of the loss distribution

$$L = -Y \quad (\text{loss})$$

that is the smallest number ℓ such that the probability that the loss L exceeds ℓ is no larger than $1 - \gamma$, for some *confidence level* $\gamma \in (0, 1)$:

$$P(L > \ell) \leq 1 - \gamma$$

(typically, $\gamma = 0.95$ or $\gamma = 0.99$)

Credit risk

Assume we create a portfolio consisting of two bonds with identical maturity T . We face the risk of a default event during the bonds' lifetimes:

- T_1, T_2 are observable lifetimes whose end is caused by the first arrival time among some unobservable shocks (*default times*)
- When $T_1 \leq T, T_2 \leq T$, our investment suffers a severe loss

Since we are interested in the joint survival probability function $P(T_1 > T, T_2 > T)$, we require additional information about the dependence between T_1 and T_2 .

Insurance risk

Let L_1, \dots, L_d , be d risks representing potential losses in different lines of business for an insurance company, which seeks protection against simultaneous big losses, in order to reduce the risk in its portfolio

One suitable reinsurance contract might be the one which pays the excess losses $L_i - k_i$ for $i \in B$, where $B = \{1, \dots, l\}$, $l \leq d$, is some set of business lines, given that $L_i > k_i$ for all $i \in B$. Hence the payout function is given by

$$f(L_i, k_i) = \left(\prod_{i \in B} \mathbb{I}_{\{L_i > k_i\}} \right) \left(\sum_{i=1}^n (L_i - k_i) \right)$$

To estimate $E(f(L_i, k_i))$ the reinsurer would typically need to study the joint distribution of (L_1, \dots, L_d) , which is however difficult, due to lack of reliable data.

Operational risk

Operational losses are usually classified in a matrix of 56 risk classes (seven event types ETs, eight business lines BLs)—see Basel Committee on Banking Supervision (2006)

Banks are required to calculate the minimum capital requirement as the 99.9%-Value-at-Risk of the loss distribution such that

$$MCR = VaR_{99.9\%} \left(\sum_{j=1}^{56} S_j \right) \quad (1)$$

where S_j is the aggregate loss of one of the 56 BL–ET combinations. It is clear that this quantity is influenced by the dependencies among the different risk classes.

1 Risk

- Definitions
- Dealing with uncertainty
- Extremes matter!

2 Loss distributions

- General Definitions
- Conditional vs Unconditional loss distribution
- Aggregate loss in Insurance

3 Final comments

In all previous examples, the choice of joint law is crucial in order to properly estimate the global risk, by taking into account the relationships among the involved variables.

The information about the dependence structure is often expressed by a single number that quantifies the degree of dependence on some scale ranging in $[-1, 1]$ (with obvious loss of information).

Moreover, in insurance and finance we often face situations where we are concerned with extreme events, where it is not the 'center of the joint distribution' that matters to us but the 'tails'

The concept of tail dependence is crucial to describe joint extreme events. Consider, for instance, two tasks:

- modeling the dependence between asset returns

The concept of **tail dependence** is crucial to describe joint extreme events. Consider, for instance, two tasks:

- modeling the dependence between asset returns
A severe crisis can cause joint drop of two (or more) stocks
- modeling the dependence between companies' default times.

The concept of **tail dependence** is crucial to describe joint extreme events. Consider, for instance, two tasks:

- modeling the dependence between asset returns
A severe crisis can cause joint drop of two (or more) stocks
- modeling the dependence between companies' default times.
Default events are rare and we want to model the probability of joint default events

In such situations, a joint distribution is needed that puts probability mass on such events and dependence cannot be describe in terms of linear association

Well-known features of the univariate Gaussian distribution are

- lack of heavy tails (implying too little probability for extreme scenarios)
- a strong form of symmetry

As a result, in many risk-management applications the *multivariate normal distribution* is not a good model since

- The joint tails of the distribution do not assign enough weight to joint extreme outcomes
- Its symmetry, known as elliptical symmetry, makes no possible to assign higher/lower probability to joint negative/positive values

Monthly log-return data

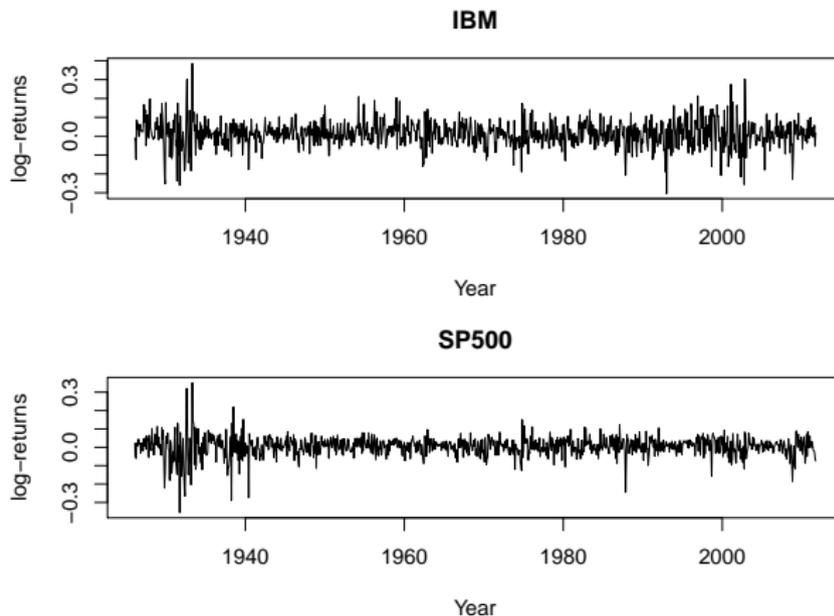


Figure: Time pots of monthly log returns of IBM stock and the S&P composite index from January 1926 to September 2011.

Motivating Example (cont)

Extremes matter!

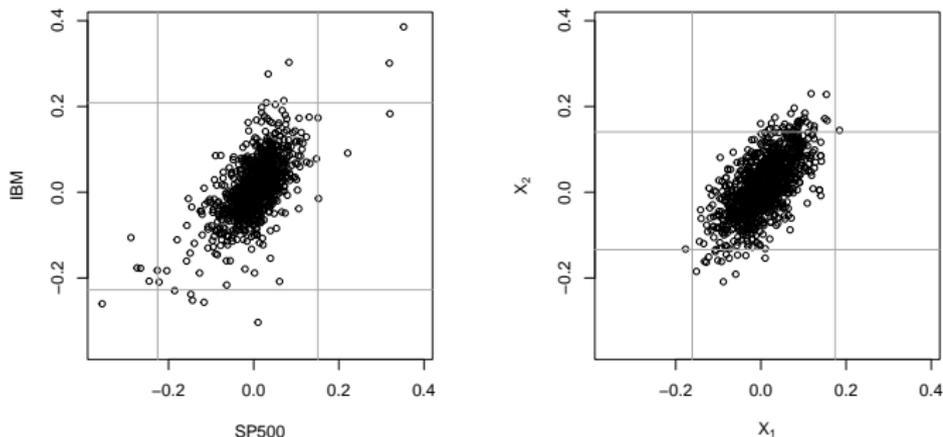


Figure: Scatterplot of SP500 and IBM Log-returns (left) and simulated data from fitted bivariate normal (right). Grey lines are quantiles at 0.005 and 0.995.

- In what follows concepts like profit-and-loss distributions, risk factors, risk measures and risk aggregation are given formal definitions and a consistent notation
- Models for risk factors need to address unexpected, abnormal or extreme outcomes, going beyond classical assumptions of Gaussianity
- The interdependence and concentration of risks is a further important challenge arising from the multivariate nature of risk

In connection with the crisis of 1998, the *Business Week* stated (September 1998)

Extreme, synchronized rises and falls in financial markets occur infrequently but they do occur. The problem with the models is that they did not assign a high enough chance of occurrence to the scenario in which many things go wrong at the same time – the “perfect storm” scenario

See also the discussion in Embrechts, Frey, McNeil (2011).

Table of Contents

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3 Final comments

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3 Final comments

In our setup, risks are random variables (rvs) on a probability space (Ω, \mathcal{F}, P) .

Risks are aggregated in a portfolio \mathcal{P} :

- ▶ a collection of financial products (stocks, bonds, derivatives, loans)

We denote with V_t the **value** of the portfolio at **time** t

Let t denote the value of \mathcal{P} today, then we are interested in the future value

$$V_{t+1}$$

of \mathcal{P} (a rv), where $t + 1$ denotes one period ahead

- ▶ e.g., 5 days (market risk), 1 year (credit and operational risk), 25 years (pension funds)

We can assume that, at time t , V_t is observable.
The **loss** L_{t+1} in time period $(t, t + 1]$ is given by

$$L_{t+1} = -\Delta V_{t+1} = -(V_{t+1} - V_t)$$

where ΔV_{t+1} is the *change* in value of the portfolio

- ▶ by convention, losses are in the right tail of the df of L_{t+1} , the **loss distribution**
- ▶ our interest is in certain statistics of the loss distribution (e.g. high quantiles)
- ▶ $(L_t)_{t \in \mathbb{N}}$ form a time series, where L_{t+1} is the loss from, say, day t to day $t + 1$

In standard risk-management practice the value V_t is modelled as a function of time and a d -dimensional random vector of *risk factors*

$$\mathbf{Z}_t = (Z_{t,1}, \dots, Z_{t,d})'$$

$$V_t = f(t, \mathbf{Z}_t)$$

where $f: \mathbb{R}_+ \times \mathbb{R}^d \rightarrow \mathbb{R}$. Examples of risk factors are (logarithmic) asset prices or exchange rates.

Risk factors are usually assumed to be observable so that, at time t , we can consider the realized value of the portfolio $v_t = f(t, \mathbf{z}_t)$.

The time series of **risk-factor changes** for one-period is $(\mathbf{X}_t)_{t \in \mathbb{N}}$, where

$$\mathbf{X}_t = \mathbf{Z}_t - \mathbf{Z}_{t-1}$$

(e.g., log-returns are changes in log prices)

Based on the risk-factor changes, we obtain the portfolio loss

$$\begin{aligned}L_{t+1} &= -(V_{t+1} - v_t) \\ &= -(f(t+1, \mathbf{Z}_{t+1}) - f(t, \mathbf{z}_t)) \\ &= -(f(t+1, \mathbf{z}_t + \mathbf{X}_{t+1}) - f(t, \mathbf{z}_t))\end{aligned}$$

Hence the loss distribution is determined by the distribution of \mathbf{X}_{t+1}

A crucial task is then finding a convenient model for L_{t+1} (or some characteristic, risk measure), based on models and statistical estimates (data) for $(\mathbf{X}_{t+1})_t$

Assume that f is differentiable. The first-order approximation of the loss (*linearized loss*) is

$$L_{t+1}^{\Delta} = - \left(f_t(t, \mathbf{z}_t) + \sum_{j=1}^d f_{z_j}(t, \mathbf{z}_t) X_{t+1,j} \right),$$

where $f_t = \partial f / \partial t$, $f_{z_j} = \partial f / \partial z_j$, for $j \in \{1, \dots, d\}$.

When risk-factor changes are likely to be small and the portfolio value is almost linear in the risk factors, it allows us to represent the loss as a linear function of the risk-factor changes.

Note that

- The quality of the approximation is best over short time horizon
- Linearization may be a problem under extremes (extreme market behavior)

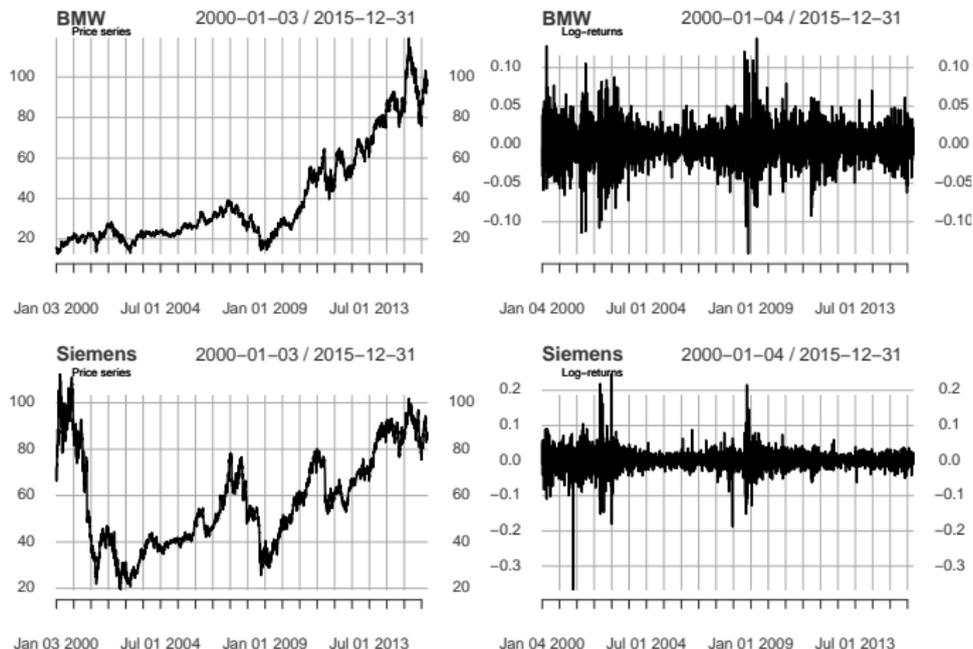


Figure: Time series of stock prices (left panels) and log-returns (right panels) for a portfolio composed of BMW, Siemens: 4174 days ($d = 2$)

Example: Portfolio of stocks (cont)

Assume \mathcal{P} consists of d stocks and let α_j ($j = 1, \dots, d$) be the number of shares of the j -th stock in \mathcal{P}

- $(S_{t,j})_{t \in \mathbb{N}}$ is the price series of the j -th stock
- Risk factors are the log transformed prices $Z_{t,j} = \ln(S_{t,j})$
- Risk-factor changes are log-returns of the stocks

$$X_{t+1,j} = Z_{t+1,j} - Z_{t,j} = \ln(S_{t+1,j}/S_{t,j})$$

Then, the value of the portfolio is

$$V_t = f(t, \mathbf{Z}_t) = \sum_{j=1}^d \alpha_j S_{t,j} = \sum_{j=1}^d \alpha_j \exp(Z_{t,j})$$

Example: Portfolio of stocks (cont)

The **loss** from time t to $t + 1$ is given by

$$\begin{aligned}L_{t+1} &= -(V_{t+1} - v_t) = -\sum_{j=1}^d \alpha_j (\exp(Z_{t+1,j}) - \exp(z_{t,j})) \\ &= -\sum_{j=1}^d \alpha_j (\exp(z_{t,j} + X_{t+1,j}) - \exp(z_{t,j})) \\ &= -\sum_{j=1}^d \alpha_j s_{t,j} (\exp(X_{t+1,j}) - 1)\end{aligned}$$

The linearized loss is given by

$$L_{t+1}^{\Delta} = -\sum_{j=1}^d \alpha_j s_{t,j} X_{t+1,j} = -v_t \sum_{j=1}^d w_{t,j} X_{t+1,j}$$

where $w_{t,j} = \alpha_j s_{t,j} / v_t$ = proportion of investment in stock j at time t

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - **Conditional vs Unconditional loss distribution**
 - Aggregate loss in Insurance
- 3 Final comments

Suppose $(\mathbf{X}_t)_{t \in \mathbb{N}}$ is a time-series of risk-factor changes.

If the distribution of $(\mathbf{X}_t)_{t \in \mathbb{N}}$ is $F_{\mathbf{X}}$ (invariant under shifts of time), then the risk-factor changes form a *stationary time series* with df $F_{\mathbf{X}}$

Let t be the current time (today). $F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$ is the conditional distribution of X_{t+1} given current information \mathcal{F}_t (*history*)

- ▶ If $(\mathbf{X}_t)_{t \in \mathbb{N}}$ is an iid series, we have $F_{\mathbf{X}_{t+1}|\mathcal{F}_t} = F_{\mathbf{X}}$

$F_{\mathbf{X}_{t+1}|\mathcal{F}_t}$ may not be equal to the stationary distribution $F_{\mathbf{X}}$

- ▶ For instance, if $(\mathbf{X}_t)_{t \in \mathbb{N}}$ are log-returns from an asset, the variance at any point in time t is often a function of past risk-factor changes and possibly of its own lagged values (e.g., GARCH models)

Conditional distribution (cont)

Given the loss L_{t+1} for the portfolio under consideration, the *conditional loss distribution* $F_{L_{t+1}|\mathcal{F}_t}$ is the distribution of L_{t+1} given all available information up to and including time t

$$F_{L_{t+1}|\mathcal{F}_t}(x) = P(L_{t+1} \leq x | \mathcal{F}_t)$$

where \mathcal{F}_t contains the historical information of the risk-factor changes up to and including time t .

Conditional distributions are particularly relevant in market risk management to study the dynamics of risk-factor changes, where the interest is in predicting volatility

Unconditional distribution

Conditional vs Unconditional loss distribution

Under the assumption that $(\mathbf{X}_t)_{t \in \mathbb{N}}$ is a stationary time series, we can consider a generic vector of risk-factor changes \mathbf{X} with the same distribution $F_{\mathbf{X}}$ as $\mathbf{X}_t, \mathbf{X}_{t-1}, \dots$, 

Hence, we neglect the modeling of dynamics and consider the *unconditional distribution* of the loss

$$F_{L_{t+1}}(x) := F_L(x) = P(L \leq x)$$

(this approach is typical in credit risk and other fields where the time horizon over which we want to measure losses is relatively large)

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3 Final comments

Consider a portfolio of n insurance contracts.

The objective is usually to build a probability model to describe the aggregate claims (number or the amount of claims) by an insurance system occurring in a fixed time period

Let S denote the aggregate losses of the portfolio in a given time period. The *individual risk model* emphasizes the loss from each individual contract and represents the aggregate losses as:

$$S_n = X_1 + \cdots + X_n$$

- the $X_i, i = 1, \dots, n$ is the loss amount from the i -th contract
- n is number of contracts in the portfolio and is a fixed number
- one usually assumes the X_i 's are independent
- the probability mass at zero corresponds to the event of no claims

A different approach considers aggregate losses in terms of a **frequency distribution** and a **severity distribution**

Let $N(t)$ the (random) number of losses over a fixed time period $[0, t]$.
If X_1, X_2, \dots are individual losses ($X_i > 0$), the *total loss amount* (*aggregate loss*) is

$$S_{N(t)} = X_1 + \dots + X_{N(t)}$$

- Each loss may or may not correspond to a unique contract (multiple claims arising from a single contract)
- Typically we assume that conditional on $N(t) = k$, the X_i 's are iid rvs
- The distribution of $N(t)$ is known as the **frequency distribution**
- the common distribution of the X_i 's is the **severity distribution**

$S_{N(t)}$ has distribution function

$$F_{S_{N(t)}}(x) = P(S_{N(t)} \leq x)$$

Whenever t is fixed (e.g. $t = 1$) we may drop the time index from the notation and simply write S_N and F_{S_N} . Some remarks:

- usually, the rvs (X_i) are iid with common df. We further assume that the rvs N and (X_i) are independent
- we may decompose the aggregate losses into the frequency process and the severity model
- the probability mass function of N is denoted by $p_{N(k)} = P(N = k), k = 0, 1, 2, \dots$
- in some cases, various forms of dependence among the X_i rvs or between N and (X_i) could be modelled

Table of Contents

- 1 Risk
 - Definitions
 - Dealing with uncertainty
 - Extremes matter!
- 2 Loss distributions
 - General Definitions
 - Conditional vs Unconditional loss distribution
 - Aggregate loss in Insurance
- 3 Final comments

- Finding the distribution of L_{t+1} is very difficult and highly dependent on the underlying risk factors.
- One thus aims at summarizing $F_{L_{t+1}}$, e.g., by means of so-called *risk measures*.
- The general task is to quantify the riskiness of a portfolio.
- If we consider the unconditional loss distribution we simply denote the df of the loss as $F_L = P(L \leq x)$, where $L = L_{t+1}, \forall t$
- Modeling dependence in finance and insurance is crucial, because it enables us to describe the relationship between variables (Example: if large losses correspond to large expenses, ignoring this relationship could lead to underestimation of required capital to cover losses)

-  Embrechts, P., Lindskog, F., & McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. Handbook of heavy tailed distributions in finance, 8(1), 329-384
-  McNeil, A. J., Frey, R., & Embrechts, P. (2015). Quantitative risk management: concepts, techniques and tools-revised edition. Princeton university press.