SAFE AI

Thorsten Schmidt

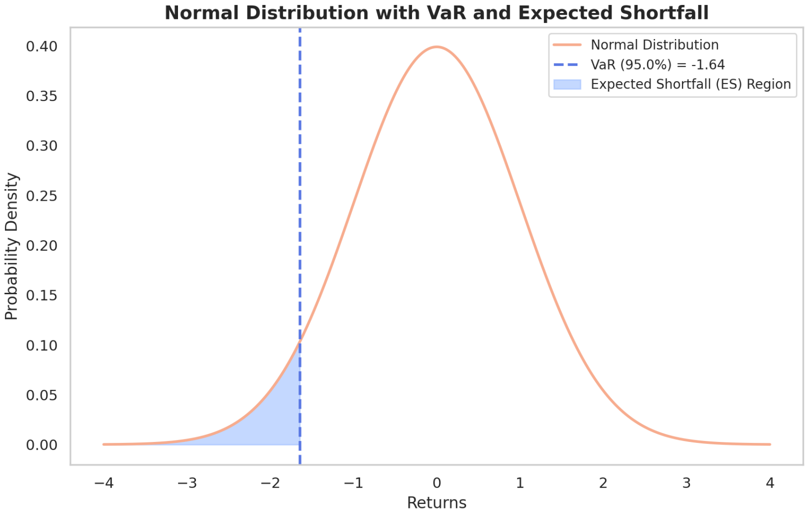
Department for Mathematical Stochastics, University Freiburg
www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

ReScale 2024

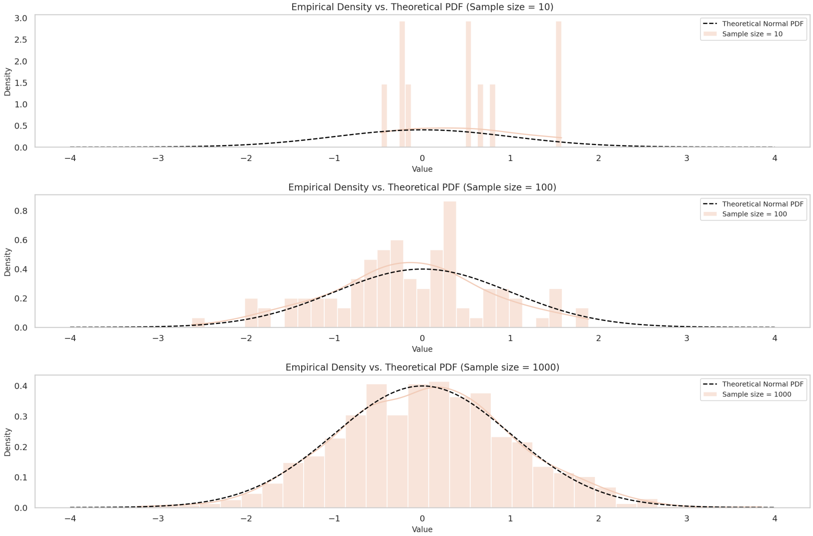
Motivation

- We give a short introduction to risk measurement of our goals and of the possible application
- risk - as we see it here - is an unwanted, unsure event in the future which (for simplicity) is associated with **monetary losses**

Risk measures



Our application



Empirical Density vs. Theoretical PDF (Sample size = 10)

Empirical Density vs. Theoretical PDF (Sample size = 100)

Empirical Density vs. Theoretical PDF (Sample size = 1000)

- The key question is how to generate the loss distribution which we will have to estimate.

Exploitation and exploration

- Now the above example simply showed some data points, but I think in our situation the case is more complex:
- I think of the kitchen application. We know what happens when we open Microwave of type 1. But there are also other types, say n_1. We know what happens when we open fridge of type 1. But there are also other types, say n_2...
- So overall we have a certain space of Elements which we call $E_1, \dots, E_N$ which contains Microwave of type 1 (element E_1), fridge of type 1 (element E_{n_1+1}) and so on.
- For some we have data (maybe small, but still) for others we do **not**
- Besides the elements there are unknowns $U_1, \dots, U_M$ which we did not test on. We have no information on these.
- Besides those we have unknown unknowns - we also have no information on the unknown unknowns.
- Our goal is to build up understanding and modelling from the bottom up.

Expert information

- I currently see two ways to enhance our datasets: transfer and experts.
- Transfer means we estimate for our robot the risk from other data sets (which we currently also do not have)
- Expert means we have a number of experts which give us information on (Simones work is the basis for this)
  - How to transfer from E_1 to $E_2, \dots, E_{n_1}$
  - and similar from the others.
  - How to asses the unknown scenarios and the associated risk
- We then follow Schmidt & Voeneky to **adaptively** gather information on the run and update on the expert estimates with incoming data in a Bayesian way.

What we need for now

- We need data on scenarios - possibly a full picture of your experiments.
- You are possibly also the experts - if we have estimates from you on how risky you estimate say the transfer from Microwave of type 1 to that of other types before you test and then test, we could gather some distribution on the quality of your estimation (and document this for later on)
- Also simulated experiments could serve as a basis for risk assessment (Joschka: autonomous cars, what are other projects where this is necessary)
- Any other ideas ?

Many thanks