UP FBE

Working Paper Series

2026/6

Máté Uzsoki, Gyöngyi Bugár, Balázs Kovács

Neural networks in developing and validating Expected Shortfall estimation models

Pécsi Tudományegyetem
Közgazdaságtudományi Kar

Neural networks in developing and validating Expected Shortfall estimation models

Máté Uzsoki, Gyöngyi Bugár and Balázs Kovács

Abstract

The new Basel III standards establish a revised framework for determining the capital charge for market risk in internal models, transitioning from Value-at-Risk (VaR) to Expected Shortfall (ES). The latter is a risk measure that more effectively captures tail risk and exhibits several advantageous properties, such as coherence. Numerous models have been proposed in the literature for estimating ES, ranging from historical simulation and parametric statistical models to quantile regression and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) methodology. Our goal is to develop a multi-level adaptive neural network which is capable to provide significantly improved ES estimates compared to a benchmark model, selected for its previous superior prediction performance compared to other models.

Keywords: Risk management, Expected shortfall, Modelling market risk, Artificial intelligence

1. Introduction and motivation

Following the outbreak of the subprime crisis in 2007 there has been a much greater need in the financial sector for valid, effective measures of risk. The new Basel III standards (BCBS, 2016) provide a revised framework for determining the capital charge for market risk in internal models with a shift from Value-at-Risk (VaR) to Expected Shortfall (ES), and this risk measure does improve the capture of tail risk. In spite of the fact that ES has gained a foothold in Basel III and is used as the proposed, new measure in estimating the risk of trading book positions, the Basel Committee still supports VaR for back-testing purposes.

It is worth mentioning that ES is not only utilized in banking practice for setting capital requirements for regulatory purposes, but that ES-based margin calculation has also spread to clearing houses (SEC, 2017; NodalClear, 2021).

However, there was a good deal of debate in the literature as to whether the validity of ES can be checked. The root of the debate was the lack of a certain property referred to as elicitability. Gneiting (2011) has pointed out that ES, in contrast to VaR, is not elicitable. It has led some voices to conclude that this risk estimator is not back-testable at all (see for instance Carver, 2013).

Due to the contribution of Acerbi and Székely (2017, 2019), the above-mentioned controversy in risk measure validation could comfortingly and convincingly be closed because there is a solution. In particular, building on its joint elicitability with Value-at-Risk (VaR), they proposed a so-called ridge back-test for ES. This development proves to be an important milestone both in absolute model validation and in model selection (relative model validation). While the former is a reality check of a single model, the latter makes it possible to choose the best-performing one from a certain variety of models.

In the literature, there are numerous proponents of different models for estimating ES. Among them are historical simulation, models based on different probability distributions (Nadarajah et al., 2015), quantile regression (Koenker, 2005) or Generalized Autoregressive Conditional Heteroscedasticity (GARCH) processes (Engle, 1982; Righi and Ceretta, 2015). A relevant question is which new modelling approaches can be used successfully in addition to already existing ones. In this area, the neural network methodology could be of interest. In recent years, several studies have demonstrated that neural network-based models can be applied with outstanding results in a number of disciplines (LeCun et al., 2015), including genetics (Leung et al., 2014), drug discovery (Ma et al., 2015), and speech recognition (Hinton et al., 2012).

Neural network (NN) methodology has been applied in the field of financial time series forecasting, mostly to predict returns (Galeshchuk, 2016; Badics, 2014; Fischer and Krauss, 2018) and to a lesser extent volatility (Lahmiri, 2017; Kristjanpoller, et al., 2014). Until the beginning of the 2020s, however, neither the neural network forecasting of the VaR nor the ES measure received special attention. The focus of these recent studies is on directly estimating VaR and/or ES using the NN-methodology

(Cheridito et al., 2020; Doncic et al., 2022; Zhang et al., 2022; Buczynski and Chlebus, 2023; Fatouros et al., 2023).

In this paper, we intend to contribute to the literature by developing a neural network that is trained with the ES-estimation error time series of a base model. The latter can be determined as desired, for instance, based on its superior performance in a former model selection. Our intention is to train the neural network so that to enhance the ES estimation of the base model.

The paper is structured as follows. Section 2 reviews the theoretical background which served as a basis for model development and presents our Artificial Neural Network – Arcsine (ANN-AS) model. Section 3 describes the data and research design. In Section 4 we discuss the results of the empirical analysis. Section 5 offers some concluding remarks.

2. Theoretical background

2.1. Value-at-Risk and Expected Shortfall as risk measures

A recent approach to quantitative risk management, as described by McNeil et al. (2015), defines portfolio risk as a statistical measure that characterises the distribution of returns (or losses) over a specified time horizon. Examples of such measures include the variance introduced by Markowitz (1952), Value-at-Risk (VaR), and Expected Shortfall (ES), with the latter two being adopted in banking regulations.

VaR is a quantile-based risk measure (Dowd and Blake, 2006), defined as the negative α-quantile of the return (X) distribution at a specified significance level α

$$VaR_\alpha(X) = -q_\alpha(X) \tag{2.1}$$

It follows from eq. (2.1)

$$VaR_\alpha(X) = -F^{\leftarrow}(\alpha), \quad 0 < \alpha < 1 \tag{2.2}$$

$F^{\leftarrow}(\alpha)$ above is the generalised inverse of the distribution function $F(x)$ for the random variable X (see Embrechts et al., 1999), i.e.

$$F^{\leftarrow}(\alpha) = inf\{x \in \mathbb{R} \,|\, F(x) \geq \alpha\}, \tag{2.3}$$

where $\inf\{.\}$ refers to the infimum, namely the greatest lower bound of the set indicated in brackets. If the α-quantile is single-valued then VaR_α can be determined as the negative inverse of the distribution function at α, namely $VaR_\alpha(X) = -F^{-1}(\alpha)$.

Since 1993, the VaR methodology has been incorporated into the banking sector's regulatory framework for determining regulatory capital needed to cover market risk-related losses. The rapid evolution of the financial sector, combined with fast-paced product innovations, continues to present

growing challenges for regulators. Events such as the subprime crisis, which escalated into a global financial crisis, highlighted the limitations of VaR as a risk measure. During this period of financial instability, banks experienced losses that significantly exceeded the regulatory capital requirements established based on VaR.

Since the early 2000s, researchers and risk professionals have increasingly criticized VaR for its limitations as a risk measure. Numerous studies have been published that emphasize the weaknesses of VaR. In 2002, the Journal of Banking and Finance even dedicated a special issue to examining issues in risk measurement theory and their applications in risk management. Szegö (2002) notably titled his editorial note with the provocative statement, 'No more VaR (this is not a typo)'. However, despite these warning signals from risk measurement theorists, banking regulators did not fully acknowledge these concerns until the onset of the subprime crisis in 2007.

Ultimately, Expected Shortfall (ES) was incorporated into Basel regulations (BCBS, 2016). At a given significance level α over a specified time horizon, ES represents a conditional expectation, specifically the expected loss (negative return) that exceeds the VaR threshold. As noted by Acerbi and Tasche (2002), it can be expressed as follows.

$$ES_\alpha(X) = VaR_\alpha(X) + \frac{1}{\alpha}E(X + VaR_\alpha)_- \tag{2.4}$$

where $(.)_- = -min(.,0)$. It is called the negative part of the given expression.

Eq. (2.4) takes the form as below.

$$ES_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha VaR_p(X)dp \tag{2.5}$$

As it is described in Bugár (2019), ES offers significant theoretical and empirical advantages over VaR. It takes into account losses that exceed VaR, which is particularly crucial for distributions with fat tails. Additionally, ES is a coherent risk measure that adheres to the coherence axioms established by Artzner et al. (1999). In contrast, VaR is not coherent due to its lack of subadditivity, which means that the risk of a portfolio, as measured by VaR, can exceed the combined risk of its individual components. Furthermore, ES has two beneficial technical characteristics: it is continuous with respect to the confidence level and convex with respect to the control variables, the latter being important for portfolio optimization. To optimize within the mean-ES framework, as demonstrated by Rockafellar and Uryasev (2000), one can solve a straightforward linear programming problem (they referred to ES as CVaR). This makes ES particularly attractive for asset allocation.

Elicitable statistics have the property of minimizing the expected value of a scoring function that quantifies the difference between forecasted and realized values from a distribution. Gneiting (2011)

pointed out that Expected Shortfall (ES) does not possess elicitability, which presents significant challenges when back-testing this risk measure.

Although back-testing has been a common practice in the financial industry and regulatory frameworks for a long time, to the best of our knowledge, its formal definition was only recently given in a contribution by Acerbi and Székely (2017).

2.2. Ridge back-test for Expected Shortfall

Acerbi and Székely (2017) noted that elicitability is a necessary condition for a risk measure to be back-testable, which implies that Expected Shortfall (ES) is not back-testable. However, in the same paper, the authors introduced the concept of a ridge back-test for ES. In a ridge back-test, the prediction of an auxiliary variable is incorporated into the back-testing process for the intended statistic, which by itself is not back-testable. This approach aligns with the minimum attained expected value of the scoring function associated with the auxiliary variable.

Based on the representation given by Rockafellar and Uryasev (2002)

$$ES_\alpha = min_{v\epsilon\mathbb{R}} E[v + 1/\alpha(X + v)_-]$$

$$VaR_\alpha = argmin_{v\epsilon\mathbb{R}} E[v + 1/\alpha(X + v)_-] \tag{2.6}$$

Acerbi and Székely (2019) have shown that ES admits a ridge back-test with the corresponding back-test function as follows:

$$Z(e, v, x) = e - v - \frac{1}{\alpha}(x + v)_- \tag{2.7}$$

where α is the given significance level, *e* and *v* denote the predicted values of ES and VaR, respectively.

The realized back-test function can be used to rank competing models by computing the mean back-test value over a period *T*, based on a series of risk measure predictions (y_t) and the corresponding realized returns (x_t)

$$\bar{z} = \frac{1}{T}\sum_{t=1}^{T} Z(y_t, x_t) \tag{2.8}$$

The test described above is directional, meaning that it can indicate a tendency for overestimation (with a positive value) or underestimation (with a negative value).

The most recent contribution by Acerbi and Székely (2019) introduces the concept of 'realized ES,' marking a notable development in the field. It can be expressed as

$$\widehat{ES}_\alpha = \frac{1}{T}\sum_{t=1}^{T} v_t + \frac{1}{\alpha}(x_t + v_t)_- \tag{2.9}$$

The average z-score can be written as the difference between the average predicted and realised value of ES

$$\bar{z}_{ES_\alpha}(e_t, v_t, x_t) = \frac{1}{T}\sum_{t=1}^{T} e_t - \frac{1}{T}\sum_{t=1}^{T} v_t + \frac{1}{\alpha}(x_t + v_t)_- \quad (2.10)$$

The importance of this finding lies in its ability to compare the predicted (ex-ante) and realized (ex-post) values of ES, ultimately allowing for what Acerbi and Székely (2019) refer to as 'apples-to-apples comparisons'.

2.3. Artificial Neural Network for enhancing ES estimation

Since McCulloch and Pitts (1943) described the first neural model, which became the basis of today's artificial neural networks (ANN), many improvements and applications of the method have appeared. The universal approximation theorems played a significant role in the spread of neural networks in recent decades. One of the earliest proofs for multilayer perceptrons (MLPs) was published by Hornik et al. (1989). It is remarkable that its first financial applications touch the field of financial forecasting. Huang et al. (2007) reviewed the applications of neural networks for financial and economic forecasting up to the mid-2000s. In their work, they identified three main areas of application, namely foreign exchange rate forecasting, forecasting stock market index price and economic growth. An overview of the literature between 2005 and 2019 is provided by Sezer et al. (2020). In the reviewed period, technological advancement enabled deep learning methods to flourish, so the first half of the study overviews the methodological background of deep learning. The authors identified areas of application such as forecasting of stock prices, indices, commodity prices, volatility, bond prices, forex prices, cryptocurrency prices and trends. The most recent results on the subject are reviewed by Zhang et al. (2023). They found that the most researched areas between 2020 and 2022 are the forecasting of stock indices, commodities, cryptocurrencies and forex rates. Their study provides an excellent overview of potential model variants, model parameters, evaluation metrics and their statistical testing, as well as implementation software environments.

Despite that neural network models for forecasting the volatility of the exchange rate existed since the mid-2000s (Sezer et al., 2020), neither the neural forecasting of the ES measure nor the VaR measure received special attention. At the beginning of the 2020s, however, several applications appeared on the subject (Buczynski and Chlebus, 2023; Cheridito et al., 2020; Doncic et al., 2022; Fatouros et al., 2023; C. X. Zhang et al., 2022). Our approach differs from the previous ones in that we do not look at the neural network as an algorithm used to estimate the ES. Instead, we determine the ES using estimated theoretical distributions at each time point and then identify the error of the estimate. With the help of the error time series, our neural network corrects the distribution-based estimates so that their expected error is minimized.

2.3.1 Description of the ANN model developed

Our ANN model utilizes a fully connected feed forward multilayer perceptron (MLP) topology (Minsky and Papert, 1969) with an input and an output layer with a single neuron each and two hidden layers with four neurons each. We used the sigmoid activation function (Wilson and Cowan, 1972) in our output and hidden layers and linear activation function in the input layer. Figure 1 shows the visual representation of the abovementioned topology with the synaptic weights and biases. The synaptic weights and biases form the set of 33 model parameters to be estimated.

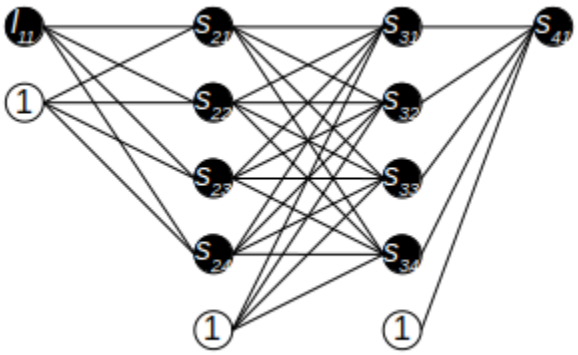


Figure 1. *The visual representation of our ANN model topology.*

In Formula 2.11 we present the algebraic formulation of the ANN model:

$$\hat{y} = f(x) = s_{41}\left(b^{41} + \sum_{i=1}^{4} w_{3i}^{41}\, s_{3i}\left(b^{3i} + \sum_{j=1}^{4} w_{2j}^{3i}\, s_{2j}\left(b^{2j} + w_{11}^{2j} l_{11}(x)\right)\right)\right) \tag{2.11}$$

where:

$\hat{y} \in [0,1]$ is the network's prediction for the output variable, normalized between 0 and 1,

f is a function that represents the relation between the input and the output of the network,

$x \in \mathbb{R}$ denotes the input variable,

s_{mn} is the sigmoid activation function of the n th neuron in the m th layer,

l_{mn} is the linear activation function of the n th neuron in the m th layer,

b^{mn} is the bias term of the n th neuron in the m th layer,

w_{kl}^{mn} is the synaptic weight coefficient between the l th neuron of the k th layer and the n th neuron of the m th layer,

i is the index of neurons in the second hidden layer,

j is the index of neurons in the first hidden layer.

The network's prediction should be denormalized to obtain the actual prediction for the dependent variable.

The goal of the network is to predict a continuous variable, so we can use the loss measures and loss minimization algorithms which are commonly used to solve regression problems in machine learning. Our learning algorithm minimizes the mean absolute error (MAE) (Wun and Pearn, 1991) using the Nadam algorithm (Dozat, 2016) for better convergence. Each of the models estimated per equity and per holding period was trained over 200 epochs in batches of 10 observations. 10 percent of the sample was retained for the validation of network training. The training sample contains the observations of *N* days preceding the given holding period of the given equity.

2.3.2. ANN-AS model

There are plenty of models proposed by the literature for ES estimation from historical simulation through parametric statistical models and quantile regression to GARCH methodology. Uzsoki (2021) has conducted a large-scale empirical investigation to compare the back-testing performance of more than 60 models. He has pointed out that among all models the arcsine distribution-based model performed best (see the Appendix for the main characteristics of this distribution). That was the reason that in this research we utilised the arcsine ES estimation model as a benchmark (base model) for training our ANN-model. From now on it is referred to as ANN-AS model.

The ANN-AS model works on the idea that it forecasts the ES-estimation error of the arcsine (AS) model for the current period ($\widehat{\delta_t}$) based on the error in the previous period (δ_{t-1}), and then corrects the ES-estimation of the AS model ($\widehat{ES}_{AS,t}$) by that value.

$$\widehat{\delta_t} = f(\delta_{t-1}) \tag{2.12}$$

$$\widehat{ES}_{ANN\text{-}AS,t} = \widehat{ES}_{AS,t} + \widehat{\delta_t} \tag{2.13}$$

where $\widehat{ES}_{ANN\text{-}AS,t}$ denotes the estimation given by the ANN-AS model in period (day) *t*.

The error of the estimation in a particular period is defined as the difference between realised ES (the term developed by Acerbi and Székely (2019)) and ES-estimation given by the AS model.

$$ESR_{\alpha,t} = v_t + \frac{1}{\alpha}(x_t + v_t)_- \tag{2.14}$$

$$\delta_t = ESR_{\alpha,t} - \widehat{ES}_{AS,t} \tag{2.15}$$

where $ESR_{\alpha,t}$ is the realised ES at significance level α in period *t*.

Please, note that according to the original definition of Acerbi and Székely (2019), the realised ES is the average of (2.14) calculated for a longer time interval *T* (normally, for a year).

The model is trained using the error time series of the AS model, by creating a data set in which the error of the previous period is coupled with the error of the current period.

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_0 & \delta_1 \\ \delta_1 & \delta_2 \\ \vdots & \vdots \\ \delta_{N-1} & \delta_N \end{bmatrix} \tag{2.16}$$

where N represents the length of the training period.

An exact description of the neural network is given in the following code fragment shown on Figure 2.

During the training, the time series of each equity was used to estimate the AS model, from which the term $ESR_{\alpha,t}$ and finally the error term was calculated.

In the phase of model development, a number of models, in all aspects significantly more complex than the one presented above, were tested, but their results were not satisfactory. The approach described above has proved to be the most appropriate for estimating ES.

```
model <- keras_model_sequential()
model %>%
   layer_dense(units = 4, activation = 'sigmoid') %>%
   layer_dense(units = 4, activation = 'sigmoid') %>%
   layer_dense(units = 1, activation = 'sigmoid')

model
  %>%
compile(loss = 'mae',optimizer = optimizer_nadam(), metrics = c('mae')  )

model %>% fit(
    X_train, Y_train,
    epochs = 200, batch_size = 10, verbose = 0,
    validation_split = 0.1
   )
   Y_predict = model %>% predict ( X_test )
```

Figure 2. *Code fragment describing the ANN-AS model.*

3. Data and research design

3.1. Database and data preprocessing

Yahoo Finance was utilized as the source of data. The empirical analysis was based on a real-data set comprising the daily closing prices of 100 equities selected from the S&P 500 constituents for the time interval of 40 years, stretching from January 1, 1983 until December 31, 2022. As there are normally 250 trading days per year, we have approximately 10,000 observations for each equity available.

In the first step of data preprocessing, we estimated the parameters of the arcsine distributions for each observed day and equity based on the returns of the 250 days preceding the given day. We then calculated VaR and ES values for the estimated distributions and assigned them to the corresponding

days. By using this data, in the next step it was possible to determine the daily components of the realized ES (see Formula 2.14). The last step was the calculation of the differences between the estimated and realized ES values, i.e., the error terms for each day and equity. The time series of error terms is used as the desired output of the neural network and by shifting this time series by one day ahead (which is the time series of error terms of the previous day) we get the input variable of the neural network.

The properties of the estimation error time series as variables used for machine learning (δ) are illustrated in Figure 3 and 4 with the data of a representative equity. Axes are not shown in Figure 3 so that all the data points can be seen. The observations are located in a very narrow band around both axes. The reason for this is the fact that most observations are close to zero. In this case, most data points are located around the origin. In the exceptional cases where VaR exceedances occur, the estimation error becomes larger by orders of magnitude. Such VaR exceedances do not usually occur on two consecutive days. For this reason, we expect most observations to have either or both coordinates close to 0. Since these points are located close to the axes, this could explain what we see in Figure 3.

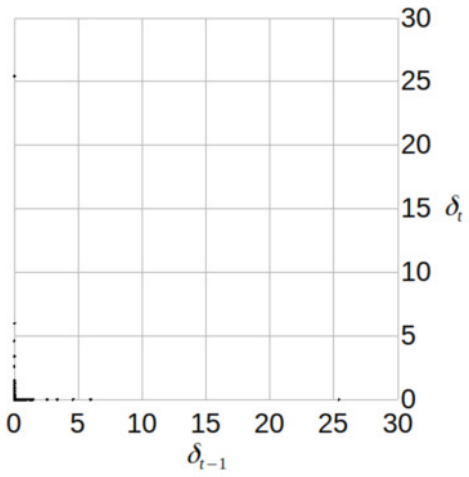


Figure 3. *Errors of the ES estimation (δ) of Apple (AAPL) using the arcsine distribution in time periods t–1 and t, in a uniformly scaled coordinate system.*

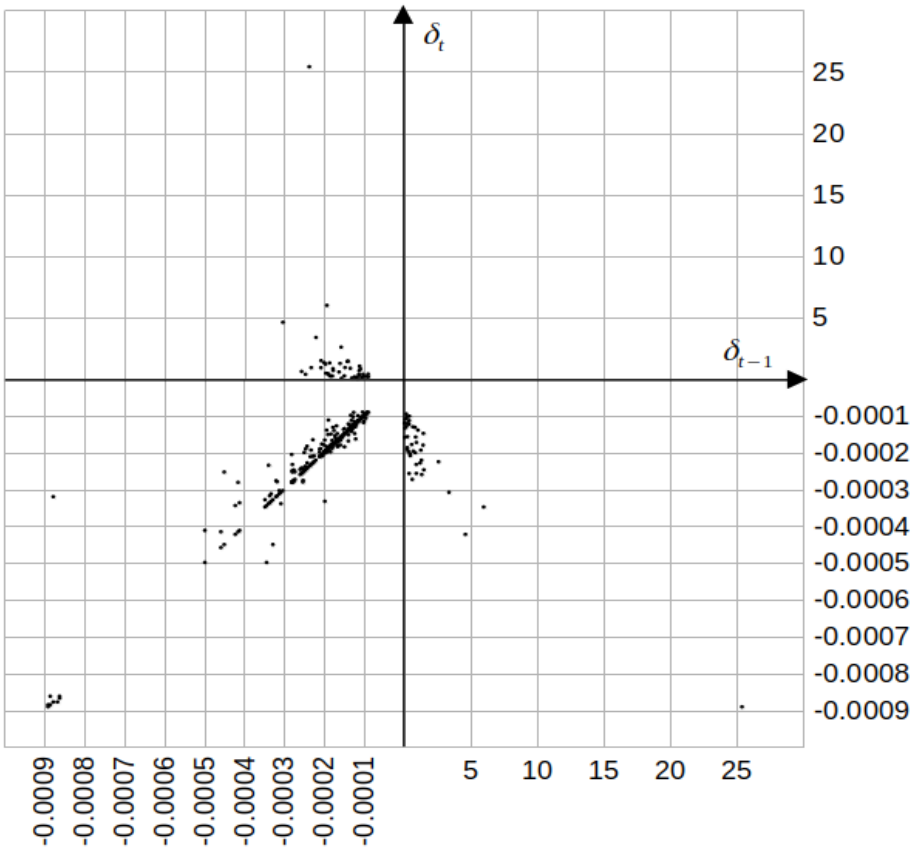


Figure 4. *Errors of the ES estimation (δ) of Apple (AAPL) using the arcsine distribution in time periods t–1 and t, in a quadrant-wise differently scaled coordinate system.*

To verify our assumption, we took a closer look at the data points around the coordinate axes, near the origin. Based on our observations, we created Figure 4, in which we applied different scaling in each quadrant of the coordinate system. Most of the observations are in the 3rd quadrant and are spreading out in the magnitude range between 0 and 10^{-3}. This is because the ES estimation by the arcsine distribution is negligibly smaller than the corresponding daily component of the realized ES (see formulas 2.14 and 2.15). In the case of the example shown in Figure 4, 0, 43, 8803 and 43 observations belong to the 1st, 2nd, 3rd and 4th quadrants, respectively. The 2nd and 4th quadrants contain those points where a VaR exceedance occurred on the current day or the previous day. In the event of a VaR exceedance, the realized ES significantly exceeds the ES value estimated from the arcsine distribution. This is the reason that in such a case we can observe orders of magnitude larger positive values. In addition to confirming our assumption, Figure 4 also provides important information about the relationship between input and output variables. In the 3rd quadrant, for example, a promising positive relationship can be identified between the estimation errors of consecutive days. Based on the small number of observations belonging to the 4th quadrant, we can assume a somewhat uncertain negative relationship. In other words, the larger error resulting from a VaR exceedance was observed the previous day, the more the arcsine model is expected to overestimate the ES value of the given day. Regarding

the 2nd quadrant, there are several arguments which are not in favor of finding an autoregressive relationship. According to the first argument, the basic assumption of risk models is that there are unpredictable individual events, and such are the VaR exceedances manifested in the output variable in this quadrant. The second argument is that, as we presented in section 2.3, neural networks are universal function approximators, and the relationships in the 3rd and 2nd quadrants cannot be modeled simultaneously with a single function, because that would violate the principle of the uniqueness of the image. The third argument is that in most cases there is no indication of a relationship between the input and the output variable in the 2nd quadrant. Using the data presented in Figures 3 and 4, for example, it can be said that the coefficient of determination of a linear model fitted to the observations of the 3rd quadrant is 99.72%, in the 4th quadrant this value is 85.76%, which is also strong, but in the 2nd quadrant, a linear model would have only 13.06% explanatory power.

Based on the insights above, we can conclude that our ANN topology with 33 model parameters is adequate to capture the changing relationship between the input and output variables over the domain.

3.2. Research design

3.2.1. Ex-ante estimation, ex-post validation

A model can be validated relying on historical data. In such a case, the sample of past returns is divided into two parts. The first part is kept for the purpose of estimation (estimation period), and the second part contains the estimators of the intended measures, which are compared to their "realised" counterparts (validation period). First, with those number of data points decided to be contained in the selected estimation period we perform an estimation, and then repeat this process several times with another estimation period – of the same length than the previous one – within the sample. The highest possible number of checkpoints can be obtained by repeating the previous process – so many times the sample makes it possible – with a one-period lag between the subsequent estimation periods.

In case of the AS model, the estimation of ES (and also that of VaR) was based on a sliding window of one-year, namely for each equity 250 data points (daily closing prices) were utilised. After the first ES (VaR) estimation has been produced, the estimation period has been moved with one day forward to get the second one, and so forth. All in all, with this rolling technique, it was possible to get an estimated ES (VaR) value for each day for the remaining 39 years comprising the validation period.

In the case of the ANN-AS model, the ES estimators were produced relying on a training period by utilising the dataset described by (2.16) with the estimation error pairs of the AS model as the input and expected output of the network. The aim was to train the network so that it could improve the ES prediction of the base model. After an appropriate network has been developed for each equity by using 4-10 years (the length of the training period) of data, a six-months holding period was set-up. The holding period referred to as the time interval, the parameters of the network has been kept unchanged, namely the network has not been modified. Therefore, for the six-month period following the training period the ES-estimators have been determined based on the same network. The network has been

retrained every six-month, and it continued repeatedly until the end of the sample period. The design of the validation process is illustrated on Figure 5.

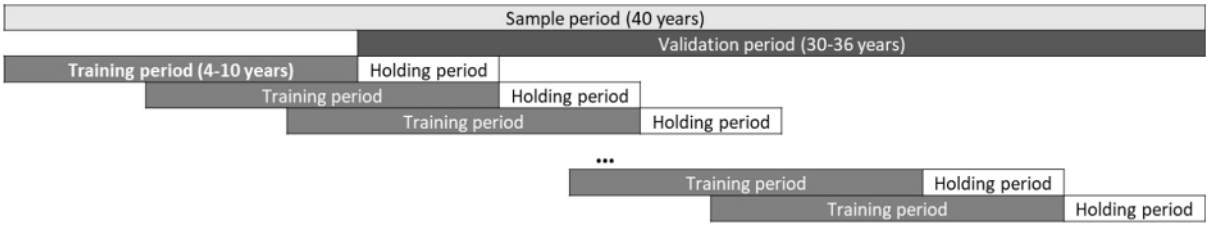


Figure 5. *The design of the validation process for the ANN-AS model.*

In our study, we present results for four different training sample sizes (N). The shortest training period consists of 4 years of data, i.e. 1000 observations. In addition to the results obtained for 5 and 6 years (1250 and 1500 observations respectively), we also examine the performance of the network estimated from 10 years', i.e. 2500 trading days' worth of data. We mentioned earlier that 33 network parameters need to be estimated, and as a rule of thumb we demanded at least 30 observations per parameter ($N \geq 999$) to ensure the reliability of the ANN model. In the cases where $N > 1500$, as we will show, the network performance increases by a negligible extent with the increase of the sample size. To represent results for larger sample sizes, we selected the 10-year long training period.

3.2.2. Performance evaluation

The forecasting performance of the ANN-AS as well as the AS models was evaluated based on the idea developed by Acerbi and Székely (2019). In particular, we calculated the Z-scores given by formula (2.7) for each and every different equity on a daily basis. The validation then has been carried out for each particular year of the validation period. The daily Z-scores have been summarised for the different years, however, rather than taking the average shown by (2.8) and (2.10), respectively.

$$Z = \sum_{p=1}^{P_y} Z_{s,y,p} \tag{3.1}$$

where $Z_{s,y,p}$ denotes the Z-score of equity s, in year y, in period p and P_y indicates the number of periods (days) in year y.

As it has been mentioned before, Z has a directional property of being able to express the tendency of over- and underestimation. However, when there is an intention to compare the overall forecasting performance of different models, an aggregation of Z-scores of different equities given by distinct years seems to be necessary. In that case, the positive and negative Z-scores might offset each other, which can result in incorrect performance evaluation, and eventually wrong conclusions. That was the reason

that for the purpose of overall performance evaluation, the absolute value of the relevant *Z*-scores has been utilised.

$$|Z|_{aggr} = \sum_{y=1}^{Y} \sum_{s=1}^{S} \left| \sum_{p=1}^{P_y} Z_{s,y,p} \right| \tag{3.2}$$

where Y denotes the number of years and S refers to the number of equities the validation has been performed.

4. Empirical results

The results of overall performance evaluation are summarised in Table 1. The aggregated absolute Z-scores have been calculated by relying on formula (3.2). Among the eight models compared there are four ANN-AS models, in particular those ones with 4 (1000), 5 (1250), 6 (1500) and 10 (2500) years (days) training period. The remaining four models include the AS model (as the baseline) and three other distribution-based models. In particular, the results are presented for assuming a t-distribution of returns with four degrees of freedom (t_4), as well as for the normal and historical cases. In the latter two, the estimation was performed by fitting a normal distribution to returns and using the empirical distribution based on historical data. The t_4, normal, and historical distributions were used as benchmarks to assess performance.

Among the different ANN-AS models considered, the ANN-AS-2500 (namely that one with a training period of 10 years) represents a bottleneck in the sense that it results in the shortest validation period of 30. Therefore, in the evaluation of all models presented in Table 1, for the purpose of comparison the last 30 years of the relevant validation period has been considered. In particular, $Y = 30$ in formula (3.2). In addition, the aggregation was based on 100 equities ($S = 100$) and 250 data points in a particular year ($P_y = 250$). In the ES-estimation a confidence (significance) level of 97.5 (2.5) has been applied.

Table 1. *Comparison of different models by aggregated absolute Z-scores.*

Model	$\|Z\|_{aggr}$
ANN-AS-2500	0.3121
ANN-AS-1000	0.3153
ANN-AS-1250	0.3154
ANN-AS-1500	0.3156
AS	0.3165
t_4	1.0241
Normal	1.0982
Historical	1.1544

As it is shown in Table 1, the ANN-AS model with 10-year training period has performed best. Furthermore, as it could have been expected, all ANN-AS models outperformed the base model (AS). The performance improvement, however, has not proved to be statistically significant.

The most striking finding is the significant difference between the aggregated absolute Z-score of the AS model and the three other distribution-based models, as shown in the last three rows of Table 1. In fact, the aggregated absolute Z-score of all the distribution-based models is more than three times higher than that of the AS model.

Further investigation is needed to understand the underlying reasons for this, but it can be noted that the heavier the tail of the distribution, the higher the ranking of the distribution-based method. For the arcsine distribution, both VaR and, consequently, ES values frequently exceed the maximum loss observed during the fitting period. Future research should explore the hypothesis that an overestimation of expected shortfall may be at play here.

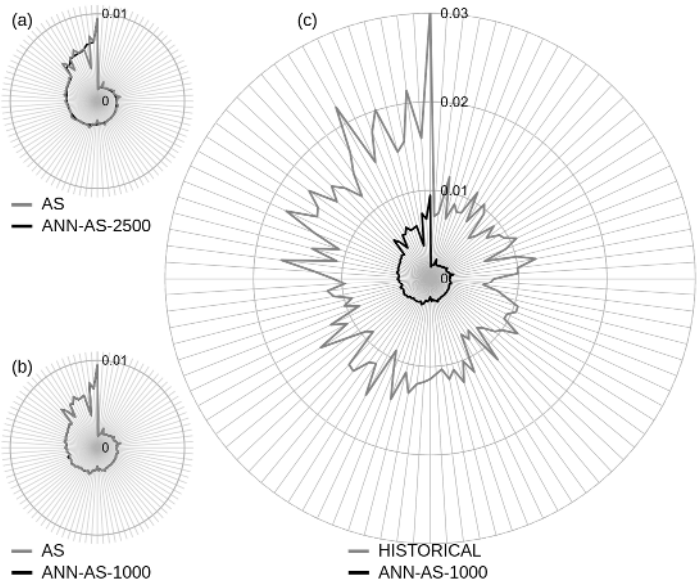


Figure 6. *Comparison of aggregated absolute Z-scores by equity-breakdown.*

Figure 6 presents a pairwise comparison of the aggregated absolute Z-scores by equity-breakdown for different models. Part (a) compares the AS and ANN-AS-2500, part (b) the AS and ANN-AS-1000, and part (c) the historical and ANN-AS-1000 models, respectively (the individual equities are not named on the figure). By summarising all relevant values plotted on the radar charts for the models considered, one can get the aggregated absolute Z-scores of the models shown in Table 1.

In 82 out of the 100 equities considered the ANN-AS-2500 model overperformed the AS model because of lower aggregated Z-score. It is remarkable that the former number is 90 in case of the ANN-AS-1000 model despite the fact that there are lower differences in the Z-scores of individual equites than in the earlier case (the graphs belonging to the two models cannot really be distinguished in part

(b) of the figure). Based on the results shown in Table 1, it is not striking that in case of all equities the ANN-AS-1000 model overperformed its historical counterpart.

Based on the results presented above, one can raise the question whether a statistically significant overperformance of the ANN-AS models to the AS model can be detected. For this purpose, for all models the yearly Z-scores defined by formula (3.1) have been considered for a 30-year validation period in case of each equity. As a result, the absolute Z-scores could be assessed for a sample size of 3000 ($= 100 \cdot 30$). Overperformance ratio is defined as a percentage of cases when the absolute Z-score of a particular model is lower than that of the AS model.

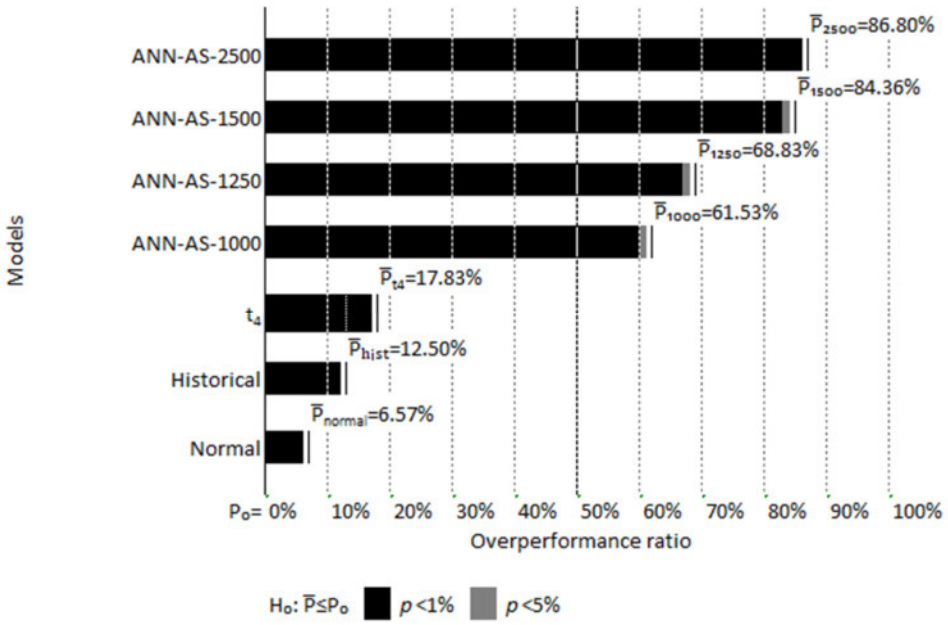


Figure 7. *Overperformance ratio of different models to the AS model.*

A basic requirement for a better-performing model is that it should produce a lower Z-score than the base model (AS) in at least 50 percent of the cases. We performed one-tailed t-tests for the performance ratios of all models with the null hypotheses $\overline{P} \leq P_0$, where $\overline{P}$ is the overperformance ratio of the corresponding model and $P_0 = 0.5$. The results of the tests reveal that while all neural networks can provide better predictions than the AS model in at least 50 percent of cases ($p \approx 0$), the historical, normal and t_4 distribution-based models are inferior in this regard ($p \approx 1$).

In practical applications, however, an overperformance ratio of more than 50 percent is usually required, so we performed our test with steps of 1 percent over the interval $P_0 \in [0,1]$. The point estimates obtained for different models can be seen in Table 2 and Figure 7.

Table 2. *Estimated overperformance ratios and their corresponding threshold values for 5 and 1 percent significance levels.*

Model	$\overline{P}$	$\max_{p<0.05} P_0$	$\max_{p<0.01} P_0$
ANN-AS-2500	86.80%	85.00%	85.00%
ANN-AS-1500	84.37%	83.00%	82.00%
ANN-AS-1250	68.83%	67.00%	66.00%
ANN-AS-1000	61.53%	60.00%	59.00%
t_4	17.83%	16.00%	16.00%
Historical	12.50%	11.00%	11.00%
Normal	6.57%	5.00%	5.00%

Based on our results, we can conclude that if we want to build a neural model that surpasses the estimation of the base model in more than 60 percent of the cases, then more than 4 years of training data are needed ($N > 1000$). With 1500 days of training data, a significantly higher performance rate of more than 80 percent can be achieved. However, the estimation advantage no longer increases proportionally by further increasing *N*. Even including additional 1000 days, only an 85 percent overperformance ratio can be guaranteed at a significance level of 1 percent.

5. Conclusion

The new back-testing method developed by Acerbi and Székely (2019) makes it possible to compare different risk estimation models and opens a way to a successful validation of Expected Shortfall as a risk measure. This plays a significant role in the Basel standards. Despite the fact that until recently the back-testability of ES has proved to be a challenge, this finding might stimulate an improvement in the "quality" of internal models as well as in the better acceptance of ES-estimation models by regulators.

In this paper, we proposed a neural network that has been trained to enhance the risk estimation performance of a base model. Considering its superior ES-estimation performance to more than 60 other models in earlier research (Uzsoki, 2021), we applied the arcsine (AS) as the base model for developing our neural network, named ANN-AS model.

We identified a non-random component in the estimation error of the realized ES, which we were able to predict with the help of our autoregressive neural network model. In our study, we found that the accuracy of the prediction can be significantly limited if the size of the training sample is chosen to be too small. According to our results, if we want to increase the proportion of predictions above 80 percent that are more accurate than the base model, then at least 6 years of training samples (actually, the fifth of validation data) must be provided to estimate the synaptic weights of the neural network. Although we have shown that an overperformance ratio exceeding 85 percent is also achievable, the 10 years of training data (the third of validation data) required for this significantly complicates the practical application of the method.

However, we do believe that improving accuracy can be achieved in other ways than increasing the sample size. We expect that resolving extreme magnitude differences in the error time series with an appropriate transformation can significantly improve network performance without the need for further increasing the sample size.

References

Acerbi, C. & Tasche, D. (2002). On the coherence of expected shortfall, *Journal of Banking and Finance*, *26*(7), 1487–1503.

Acerbi, C. & Székely, B. (2017). General Properties of Backtestable Statistics. *Working paper.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2905109.

Acerbi, C. & Székely, B. (2019). *The minimally biased backtest for ES*. Risk.net, 29 August. https://www.risk.net/cutting-edge/banking/6947266/the-minimally-biased-backtest-for-es.

Artzner, P., Delbaen, F., Eber, J. M. & Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, *9*, 203-228.

Badics, M. C. (2014). Tőzsdei idősorok előrejelzése adatbányászati módszerekkel / Forecasting stock exchange time series with data mining methods. *Hitelintézeti Szemle*, *13*(4), 207-227.

BCBS (2016). Basel Committee in Banking Supervision *Minimum Capital Requirements for Market Risk.* Bank for International Settlements, 1–88.

Buczynski, M. & Chlebus, M. (2023). GARCHNet: Value-at-Risk Forecasting with GARCH Models Based on Neural Networks. *Computational Economics*. https://doi.org/10.1007/s10614-023-10390-7.

Bugár, G. (2019). A Breakthrough Idea in Risk Measure Validation – Is the Way Paved for an Effective Expected Shortfall Backtest?, Financial and Economic Review, *18*(4), 130-145.

Carver, L. (2013). Mooted VAR substitute cannot be back-tested, says top quant. *Risk, 8*, March. https://www.risk.net/regulation/basel-committee/2253463/mooted-var-substitute-cannot-be-back-tested-says-top-quant.

Cheridito, P., Ery, J. & Wüthrich, M. V. (2020) Assessing asset-liability risk with neural networks. *Risks*, *8*(1). https://doi.org/10.3390/risks8010016.

Doncic, S., Pantić, N., Lakicević, M. & Radivojević, N. (2022). Expected shortfall model based on a neural network. *Journal of Risk Model Validation*, *16*(2). https://doi.org/10.21314/JRMV.2022.016.

Dowd, K. & Blake, D. (2006). After VaR: The Theory, Estimation, and Insurance Applications of Quantile Based Risk Measures, *The Journal of Risk and Insurance, 73*(2), 193-229.

Dozat, T. (2016). Incorporating Nesterov Momentum into Adam. *ICLR Workshop,* 2-4 May 2016.

Embrechts, P., Resnick, S.I. & Samorodnitsky, G. (1999). Extreme Value Theory as a Risk Management Tool, *North American Actuarial Journal*, *3*(2), 30–41.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987-1007.

Erdős, P. & Kac, M. (1947). On the number of positive sums of independent random variables. *Bulletin of the American Mathematical Society*, *53*(10), 1011–1020.

Fatouros, G., Makridis, G., Kotios, D., Soldatos, J., Filippakis, M. & Kyriazis, D. (2023). DeepVaR: a framework for portfolio risk assessment leveraging probabilistic deep neural networks. *Digital Finance*, *5*(1), 29-56. https://doi.org/10.1007/s42521-022-00050-0.

Fischer, T. & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669.

Galeshchuk, S. (2016). Neural networks performance in exchange rate prediction. *Neurocomputing*, *172*, 446-452.

Gneiting, T. (2011). Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, *106*, 746-762.

Hinton, G., Deng, L., Yu, D., Dahl, G. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P. Sainath, T. & Kingsbury B. (2012). Deep neural networks for acoustic modeling in speech. *IEEE Signal Processing Magazine*, *29*, 82-97.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang, W., Lai, K. K., Nakamori, Y., Wang, S. & Yu, L. (2007). Neural networks in finance and economics forecasting. *International Journal of Information Technology and Decision Making*, *6*(1), 113-140. https://doi.org/10.1142/S021962200700237X.

Koenker, R. (2005). *Quantile Regression.* Cambridge: Cambridge University Press.

Kristjanpoller, W., Fadic, A. & Minutolo, M. C. (2014). Volatility forecast using hybrid Neural Network models. *Expert Systems with Applications*, *41*, 2437-2442.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444.

Lahmiri, S. (2017). Modeling and predicting historical volatility in exchange rate. *Physica*, 387-395.

Leung, M., Xiong, H., Lee, L. & Frey, B. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics*, *30*, 121-129.

Ma, J., Sheridan, R. P., Liaw, A. Dahl, G. E. & Svetnik V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, *55*(2), 263-274.

Markowitz, H. M. (1952). Portfolio selection, *Journal of Finance*, *8*, 77-91.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259.

McNeil, A. J., Frey, R. & Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, New Jersey.

Minsky, M. L. & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Nadarajah, S., Chan, S. & Afuecheta, E. (2015). *Package 'VaRES'*. https://cran.r-project.org/web/packages/VaRES/VaRES.pdf

NodalClear (2021). *Margin Methodology*. https://www.nodalexchange.com/nodal-clear/risk-management/margin-methodology/

Righi, M. B. & Ceretta, P. S. (2015). A comparison of Expected Shortfall estimation models. *Journal of Economics and Business*, *78*, 14-47.

Rockafellar, R.T. & Uryasev, S. (2000). Optimization of Conditional Value-at-Risk, *Journal of Risk*, *3*(3), 21-41.

Rockafellar, R.T. & Uryasev, S. (2002). Conditional Value-at-Risk for General Loss Distributions, *Journal of Banking and Finance*, *26*, 1443-1471.

SEC (2017). *Release No. 34-82313* (File No. SR-ICEEU-2017-013). https://www.sec.gov/rules/sro/iceeu/2017/34-82313.pdf.

Sezer, O. B., Gudelek, M. U. & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing Journal*, *90*. https://doi.org/10.1016/j.asoc.2020.106181.

Szegö, G. (2002). No more VaR (this is not a typo), *Journal of Banking and Finance*, *26*, 1247-1251.

Uzsoki, M. (2020). Neurális hálók a pénzügyi kockázatmérésben / Neural Networks in Measuring Financial Risk, *Szigma*, *51*(3), 287-299.

Uzsoki, M. (2021). A piaci kockázat mérése neurális hálózattal és statisztikai modellekkel / Measuring Market Risk with Neural Network and Statistical Models, *PhD dissertation*, University of Pécs.

Wilson, H. R. & Cowan, J. D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, *12*(1), 1-24. https://doi.org/10.1016/S0006-3495(72)86068-5.

Wun, L. M. & Pearn, W. L. (1991). Assessing the statistical characteristics of the mean absolute error or forecasting. *International Journal of Forecasting*, *7*(3), 335-337. https://doi.org/10.1016/0169-2070(91)90007-I.

Zhang, C., Sjarif, N. N. A. & Ibrahim, R. B. (2023). Deep learning techniques for financial time series forecasting: A review of recent advancements: 2020-2022. *Working paper*. https://arxiv.org/ftp/arxiv/papers/2305/2305.04811.pdf.

Zhang, C. X., Li, J., Huang, X. F., Zhang, J. S. & Huang, H. C. (2022). Forecasting stock volatility and value-at-risk based on temporal convolutional networks. *Expert Systems with Applications*, *207*. https://doi.org/10.1016/j.eswa.2022.117951.

Appendix

Main characteristics of the arcsine distribution

The arcsine distribution is a special case of the beta distribution: $X \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$, known from the works of Erdős and Kac (1947).

The probability density function of the arcsine distribution can be described as:

$$f(x) = \frac{1}{\pi\sqrt{(x-a)(b-x)}}$$

The cumulative distribution function of the arcsine distribution is defined by the following formula:

$$F(x) = \frac{2}{\pi}\arcsin\left(\sqrt{\frac{x-a}{b-a}}\right)$$

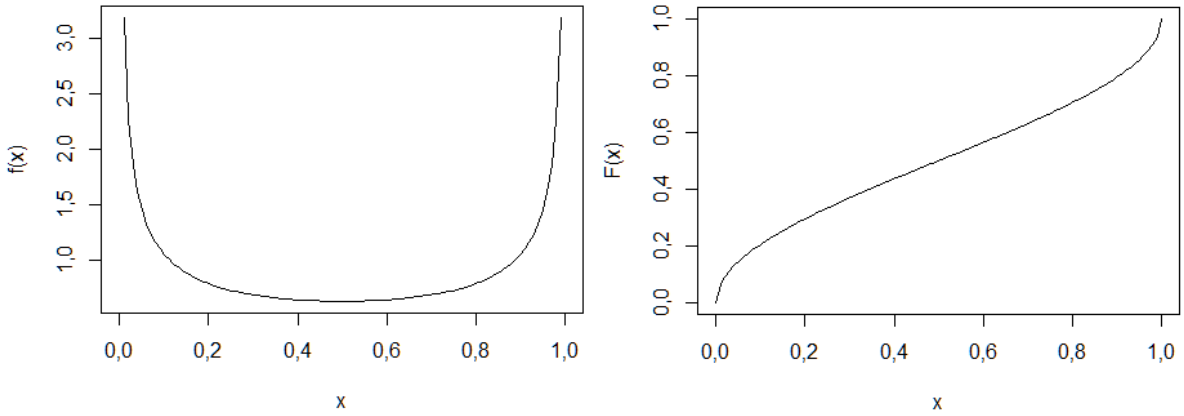


Figure 8. *Density function and cumulative distribution function of the arcsine distribution with standard parameters*

The value at risk and expected shortfall of the distribution are as follows (Nadarajah et al., 2015).

$$VaR_p(X) = a + (b-a)\sin^2\frac{\pi p}{2}$$

$$ES_p(X) = a + \frac{b-a}{p}\int_0^p \sin^2\frac{\pi v}{2}\,dv$$

where:

$a < x < b$

$0 < p < 1$

$-\infty < a < b < \infty$.