Case Study 2

1. Calibrating a **kernel PDF** on empirical data
2. Testing the compatibility of the calibrated distribution with respect to original data
3. BONUS: generating random numbers from a target distribution
4. Applications to risk management

Kernel density

1. The main idea is to find an (almost) **assumption-free** PDF to fit some data

2. Typically kernel densities are given by the sum of local densities centred around a portion of the available data, which we shall refer to as the **“training set”**

3. In addition, we also need another portion of the data to estimate the parameters of the local densities (**“validation set”**), and possibly yet another portion to test the out-of-sample performance of the calibrated density (**“testing set"**)

Case Study 2 - Useful formulas

General Kernel

$$p(x; h, \{x_i\}_{i \in T}) = \frac{1}{N_T h} \sum_{i \in T} K\left(\frac{x - x_i}{h}\right)$$

Point at which the full PDF is computed

Fixed set of data points ("training set")

Size of training set

Bandwidth parameter

Case Study 2 - Useful formulas

General Kernel

$$p(x; h, \{x_i\}_{i \in T}) = \frac{1}{N_T h} \sum_{i \in T} K\left(\frac{x - x_i}{h}\right)$$

Assuming a standard Gaussian kernel, i.e. $K \to N(0, 1)$ we get:

Gaussian Kernel

$$p(x; h, \{x_i\}_{i \in T}) = \frac{1}{N_T \sqrt{2\pi h^2}} \sum_{i \in T} \exp\left(-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right)$$

CDF of Gaussian Kernel

$$C(x; h, \{x_i\}_{i \in T}) = \frac{1}{2N_T} \sum_{i \in T} \left[1 + \operatorname{erf}\left(\frac{x - x_i}{\sqrt{2}h}\right)\right]$$

CCDF of Gaussian Kernel

$$C(x; h, \{x_i\}_{i \in T}) = \frac{1}{2N_T} \sum_{i \in T} \left[1 - \operatorname{erf}\left(\frac{x - x_i}{\sqrt{2}h}\right)\right]$$

Maximum likelihood

Likelihood function

$$\mathcal{L}(h) = \prod_{j \in V} p(x_j; h; \{x_i\}_{i \in T})$$

"Validation set"

$$V \cap T = 0$$

Fixed set of data points ("training set")

Maximum likelihood

Likelihood function

$$\mathcal{L}(h) = \prod_{j \in V} p(x_j; h; \{x_i\}_{i \in T})$$

log-likelihood function

$$\log \mathcal{L}(h) = \sum_{j \in V} \log p(x_j; h, \{x_i\}_{i \in T})$$

Maximum likelihood problem

$$h_{\text{opt}} = \underset{h}{\text{argmax}} \ \log \mathcal{L}(h)$$

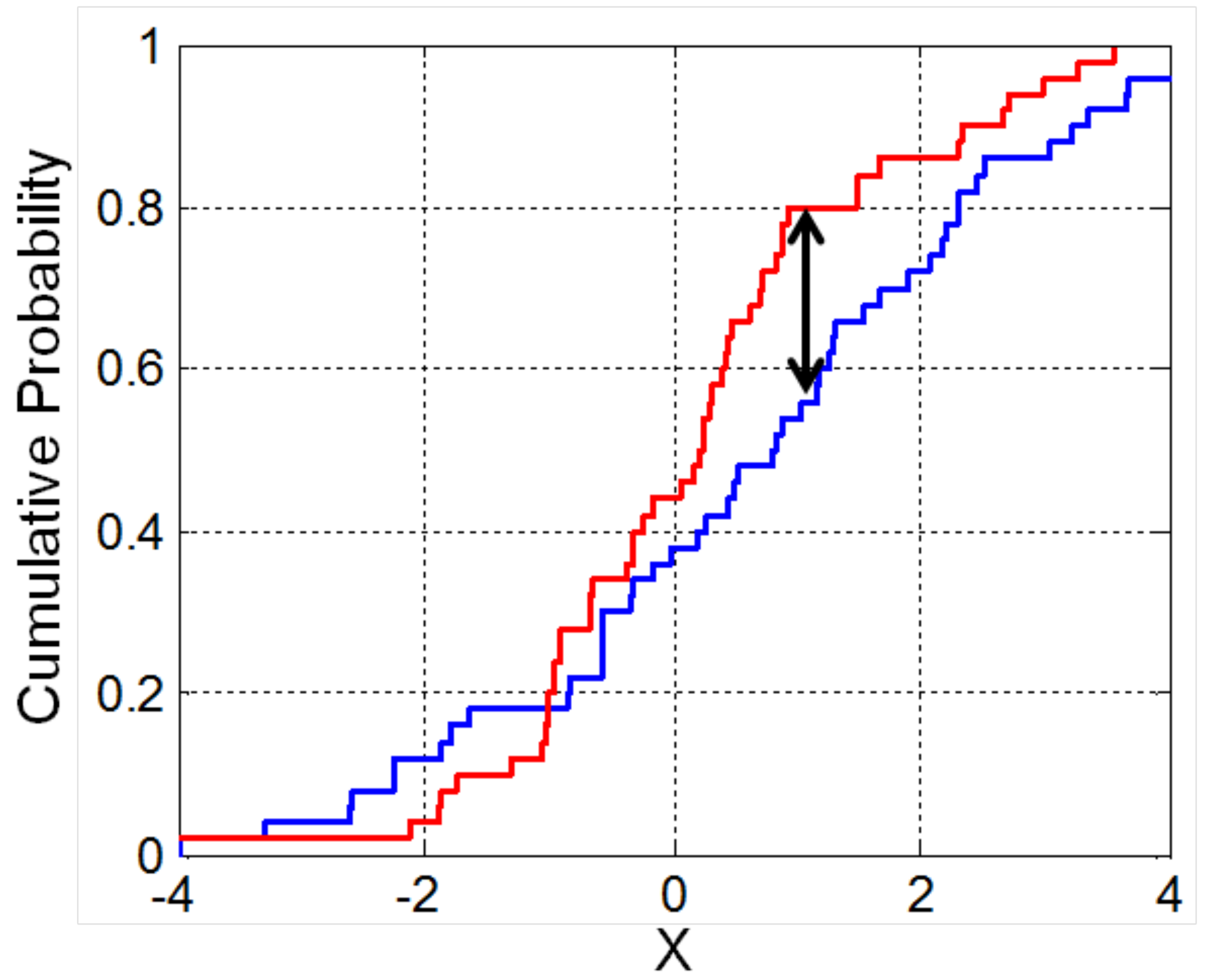
Optimal (most likely) bandwidth

Two sided Kolmogorov-Smirnov test

1. Testing the null hypothesis that two data samples (possibly of different size) are generated by the same distribution

2. This is done by quantifying the probability of the largest observed distance between the empirical cumulative distributions of the two data samples

$$D_{nm} = \sup_x |C_{1,n}(x) - C_{2,m}(x)|$$



Random number generation via inversion of CDF

1. By construction cumulative distribution functions (CDFs) produce as output numbers in [0,1]

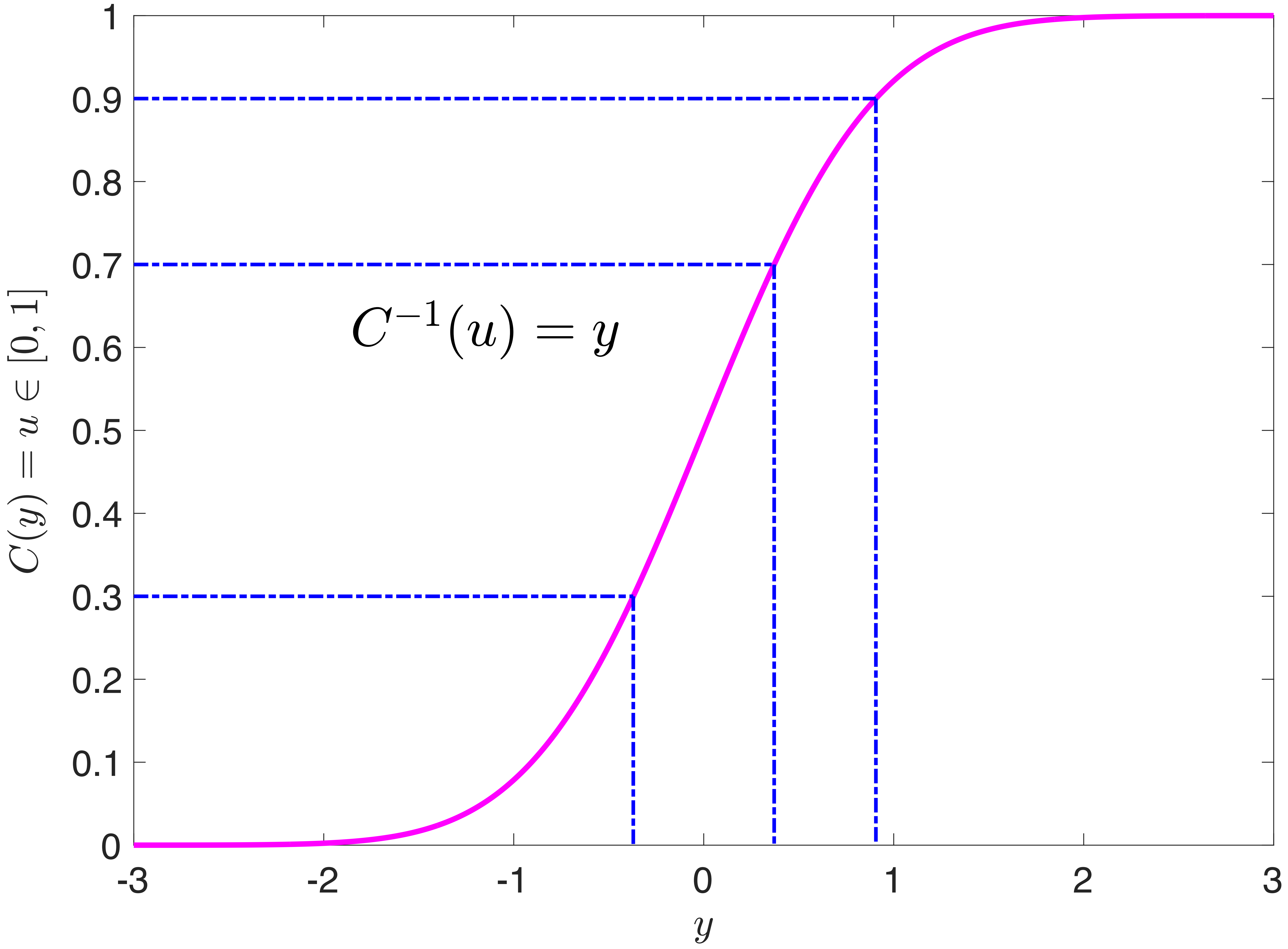
$$C(y) = \int_{-\infty}^{y} p(x)\ dx = u \in [0, 1]$$

2. This can be exploited to generate random numbers from a desired target distribution

3. This can be done by drawing random numbers from the uniform distribution in [0,1] and mapping them to the desired distribution by inverting the CDF

$$u \in [0, 1] \implies C^{-1}(u) = y$$

Random number generation via inversion of CDF



Value-at-Risk and Expected Shortfall

1. At a certain significance level α, the Value-at-Risk is the (1-α) quantile of the return distribution

$$1 - \alpha = \int_{-\infty}^{-\mathrm{VaR}_\alpha} p(r)\ dr$$

2. The Expected Shortfall is the expected loss in the α% worst cases (average loss worse than the VaR)

$$\mathrm{ES}_\alpha = -\tfrac{1}{1-\alpha} \int_{-\infty}^{-\mathrm{VaR}_\alpha} r\ p(r)\ dr$$