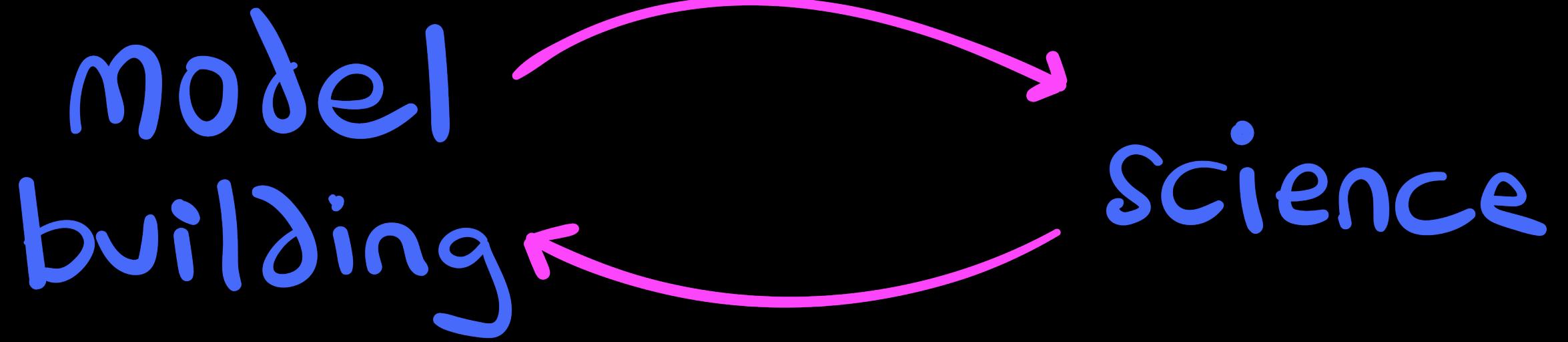


6. Connecting theory to practice: a first look at model building



6.1. Data and random samples

S



$\rightarrow R$

$X = \text{price}$

(listing 1, listing 2, listing 3, ..., listing n)

$x_1 \downarrow$

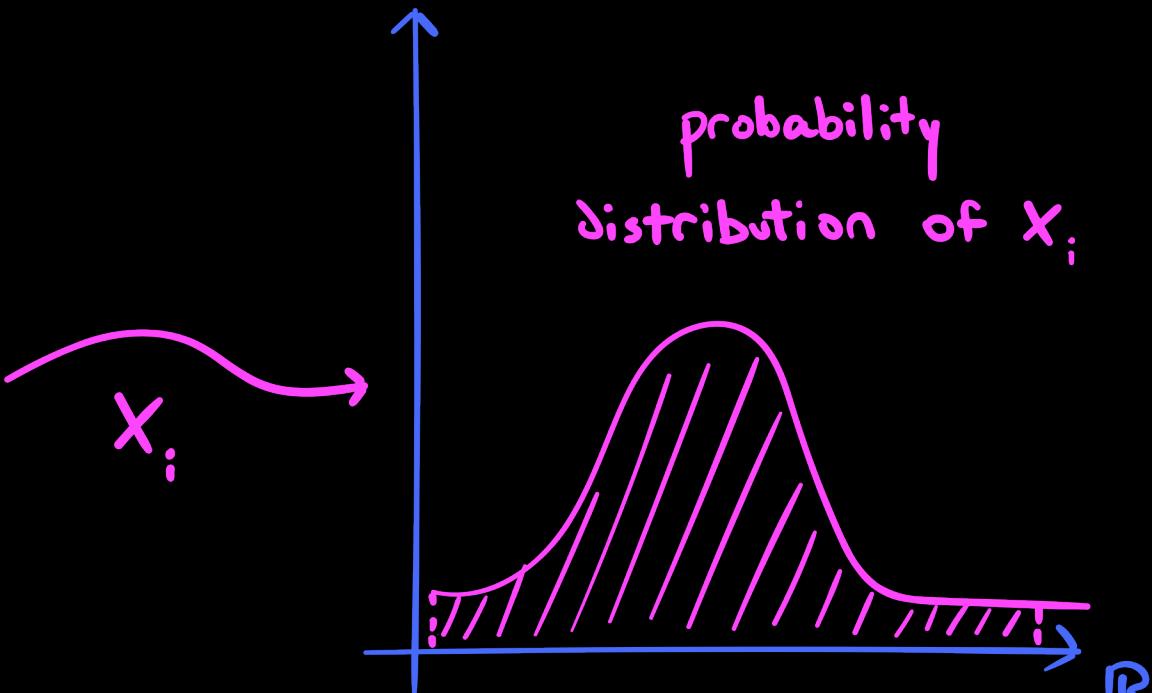
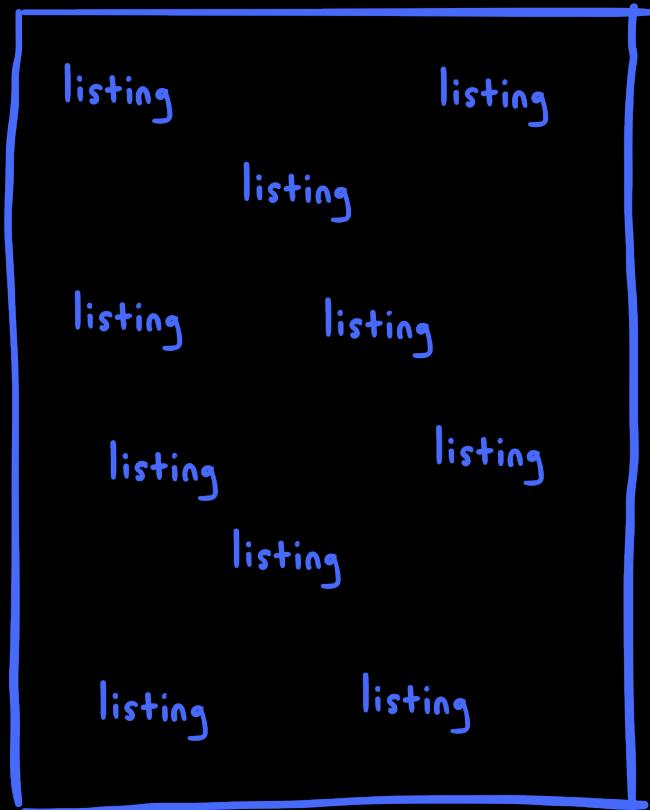
$x_2 \downarrow$

$x_3 \downarrow$

$x_n \downarrow$

(price 1, price 2, price 3, ..., price n)

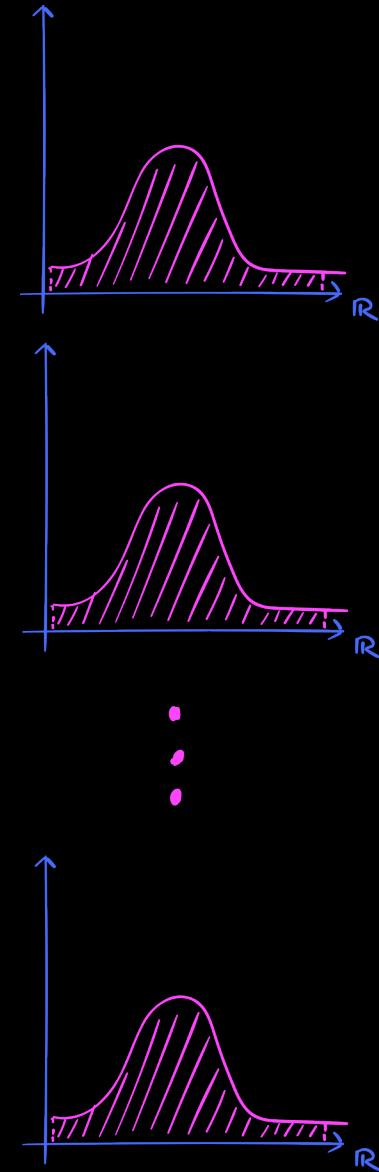
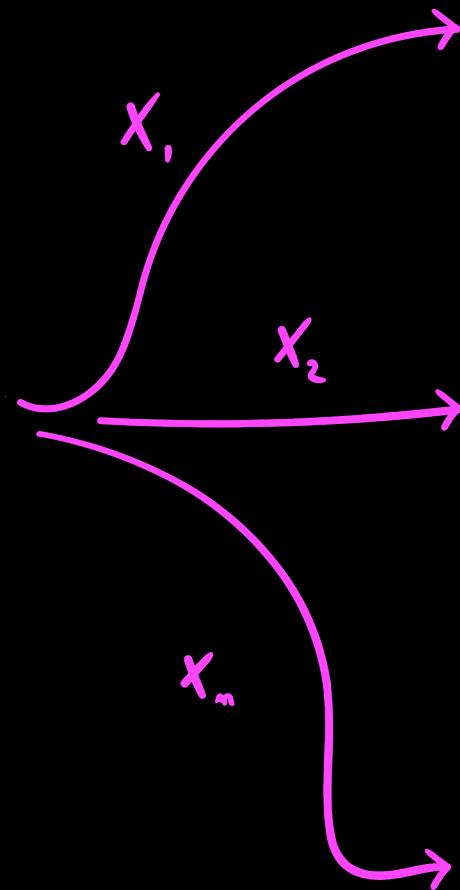
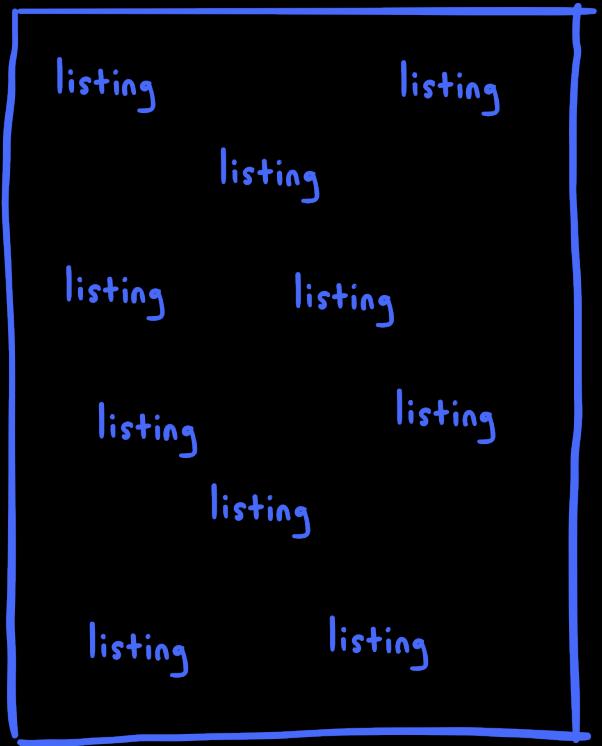
S



these prices
are common...

...these ones
less so

S



identical
distributions!!!

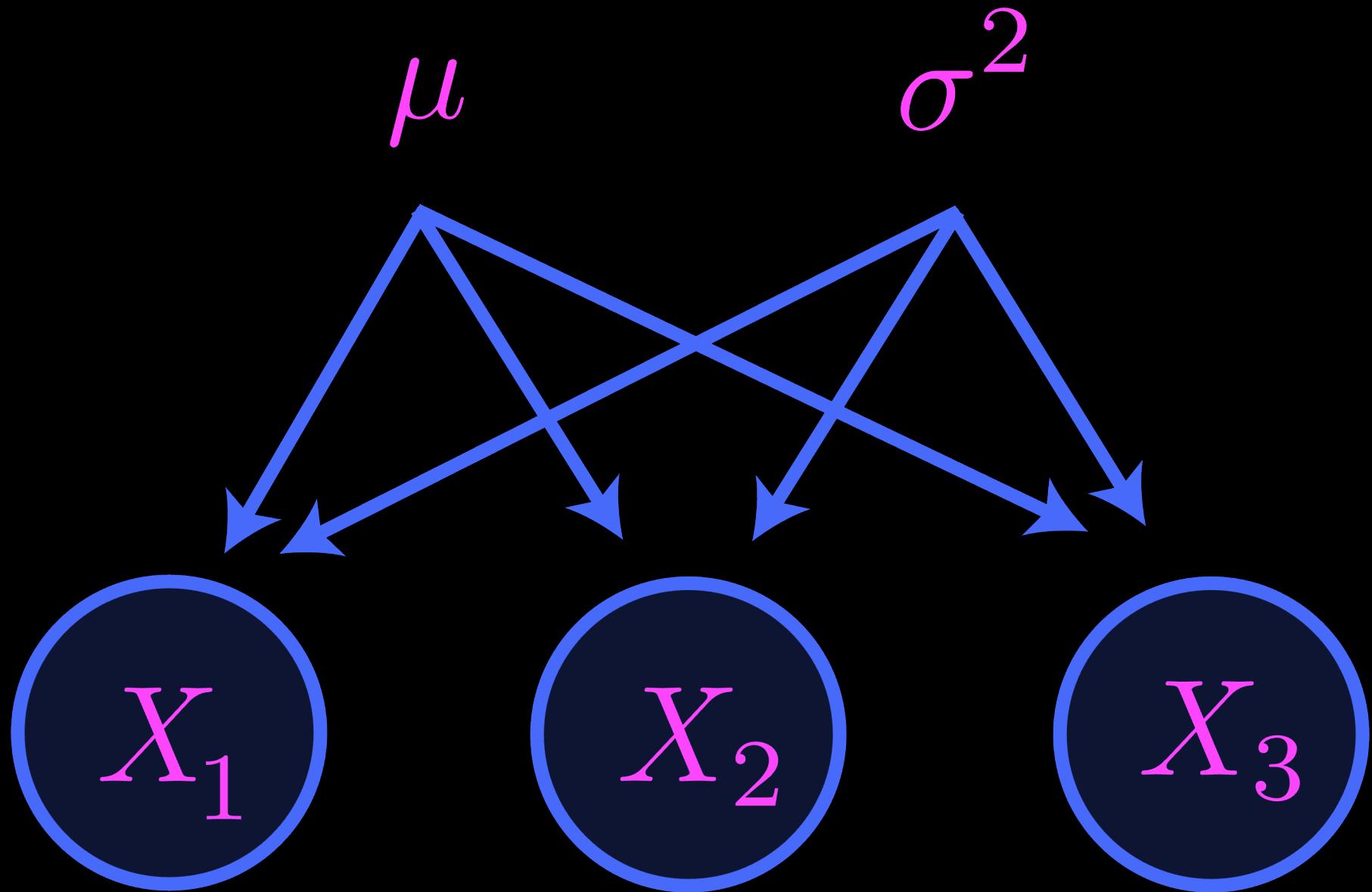
 **Definition 6.1**

Let X_1, X_2, \dots, X_m be a sequence of random variables, all defined on the same probability space.

- The random variables are called a *random sample* if they are *independent* and *identically distributed* (IID).

Provided that the sequence is a random sample, an *observed random sample*, or a *dataset*, is a sequence of real numbers x_1, x_2, \dots, x_m where x_i is an observed value of X_i . We shall also refer to a dataset as an *observation* of the corresponding random sample.

6.2. Probabilistic models and empirical distributions



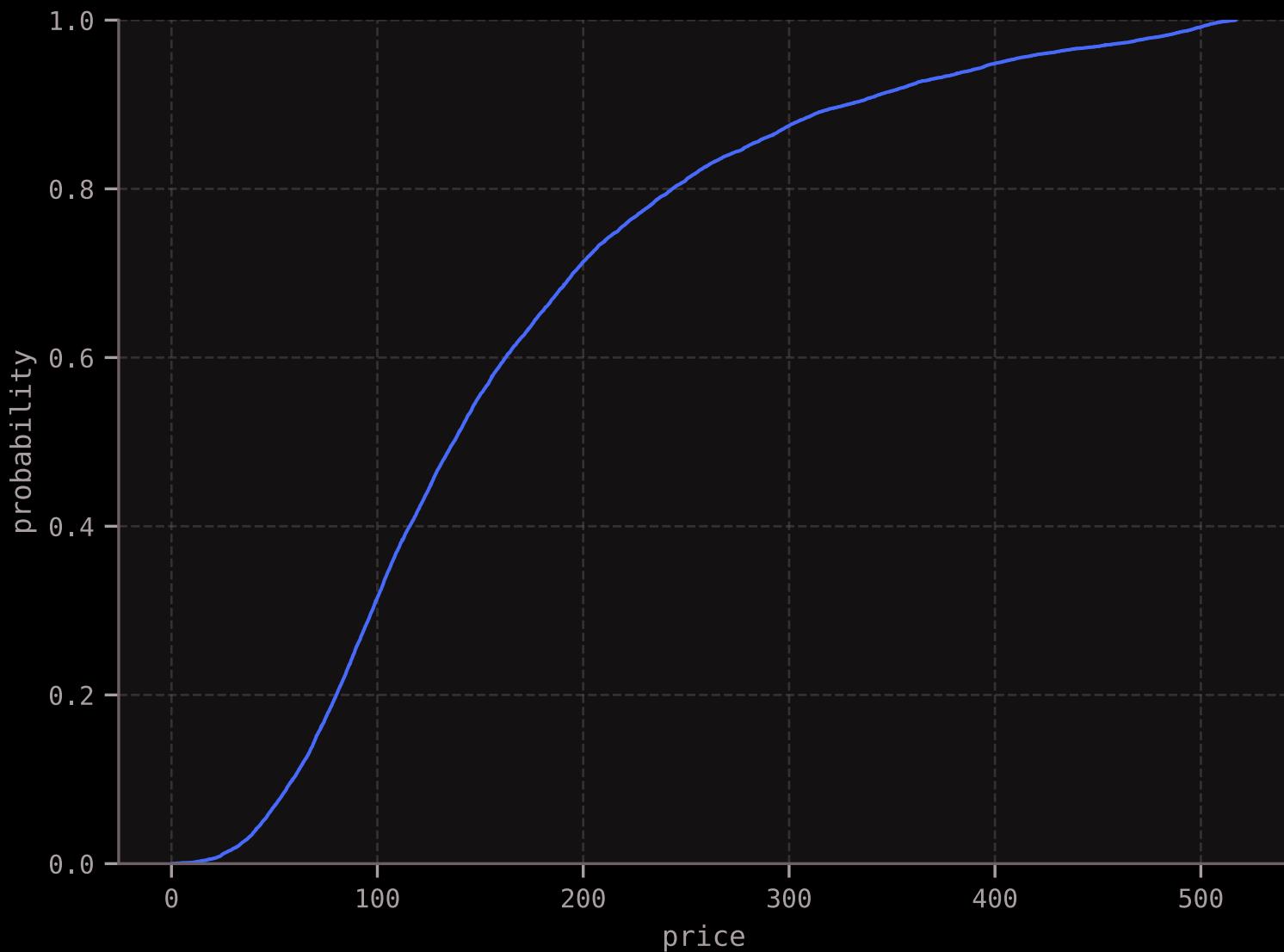
Definition 6.2

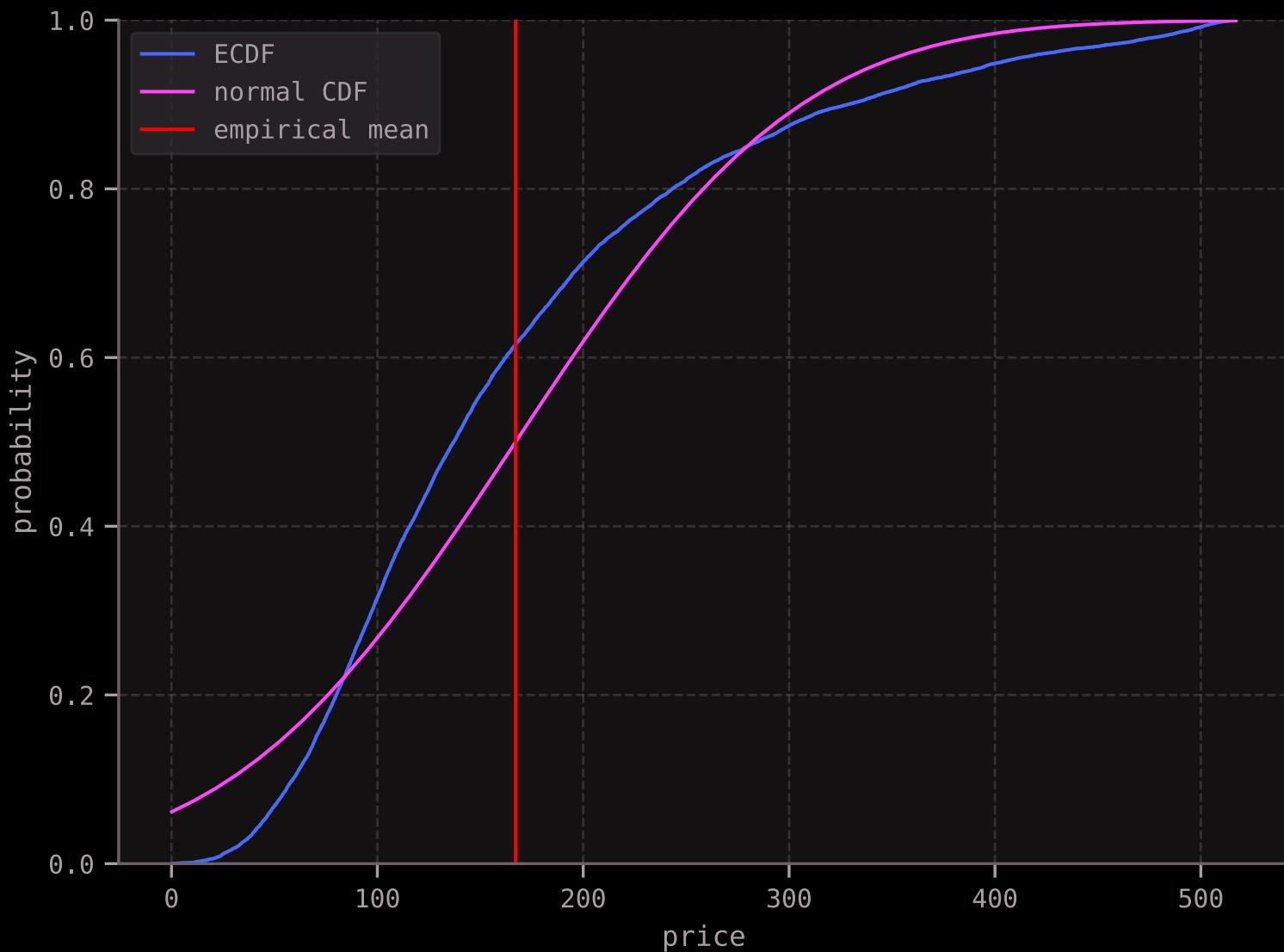
The *empirical distribution* of a dataset x_1, x_2, \dots, x_m is the discrete probability measure on \mathbb{R} with probability mass function

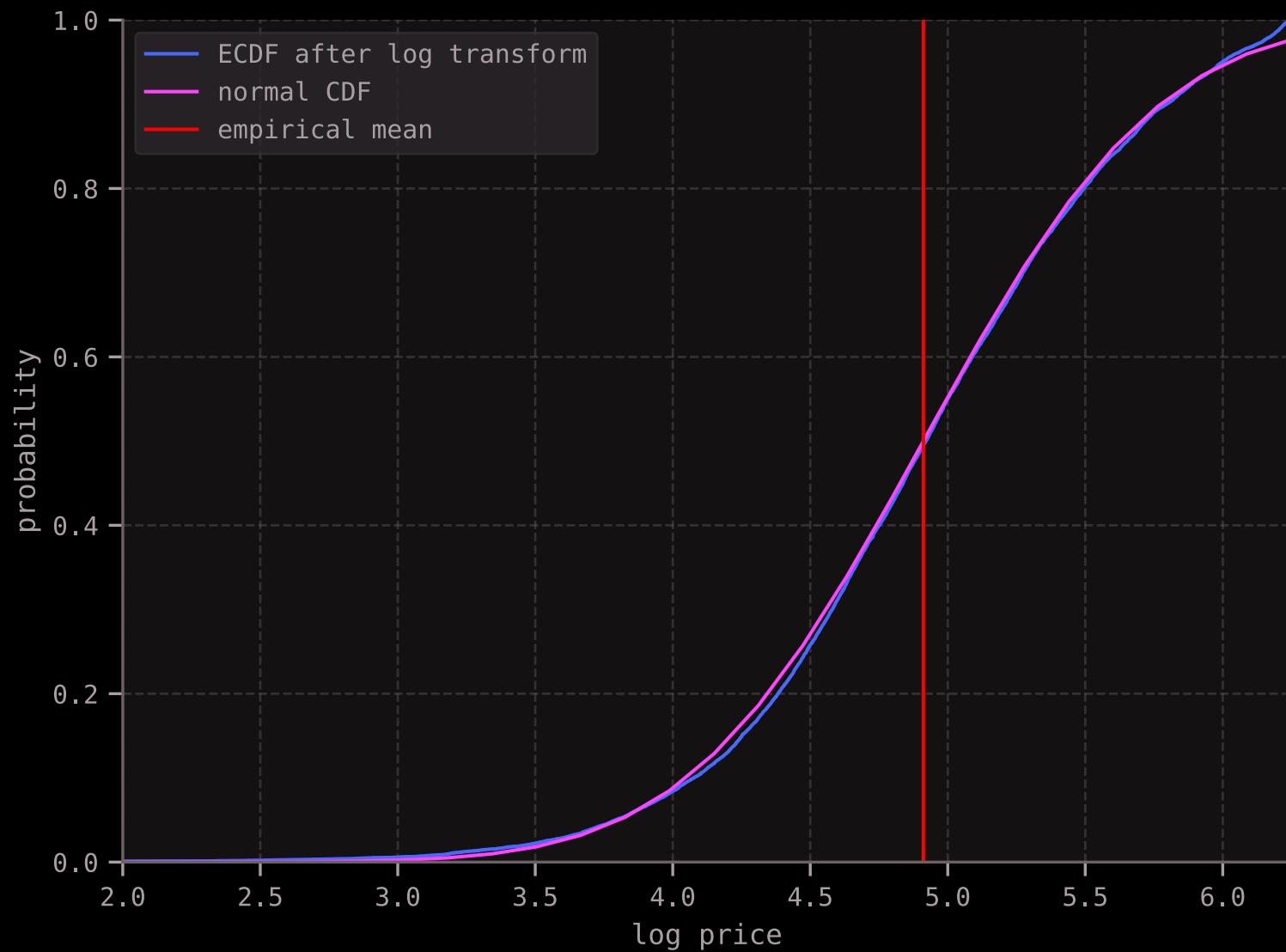
$$p(x) = \frac{\text{number of data points } x_i \text{ that match } x}{m}.$$

The *empirical cumulative distribution function (ECDF)* of the dataset is the CDF of the empirical distribution. It is given by

$$F(x) = \sum_{x^* \leq x} p(x^*) = \frac{\text{number of data points } x_i \text{ with } x_i \leq x}{m}.$$





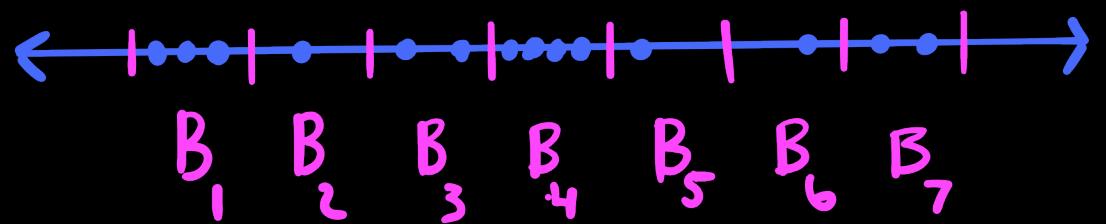


6.3. Histograms

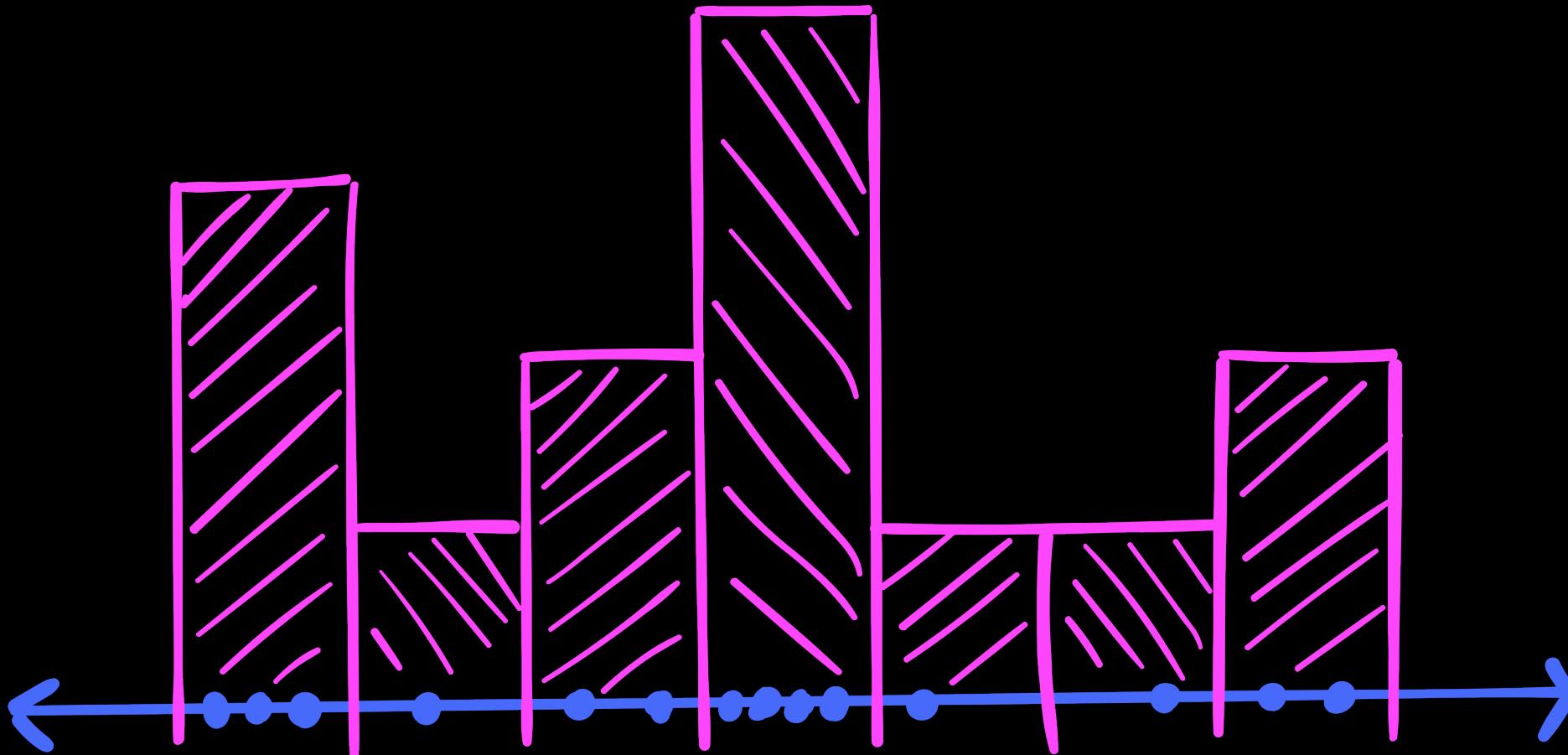
dataset

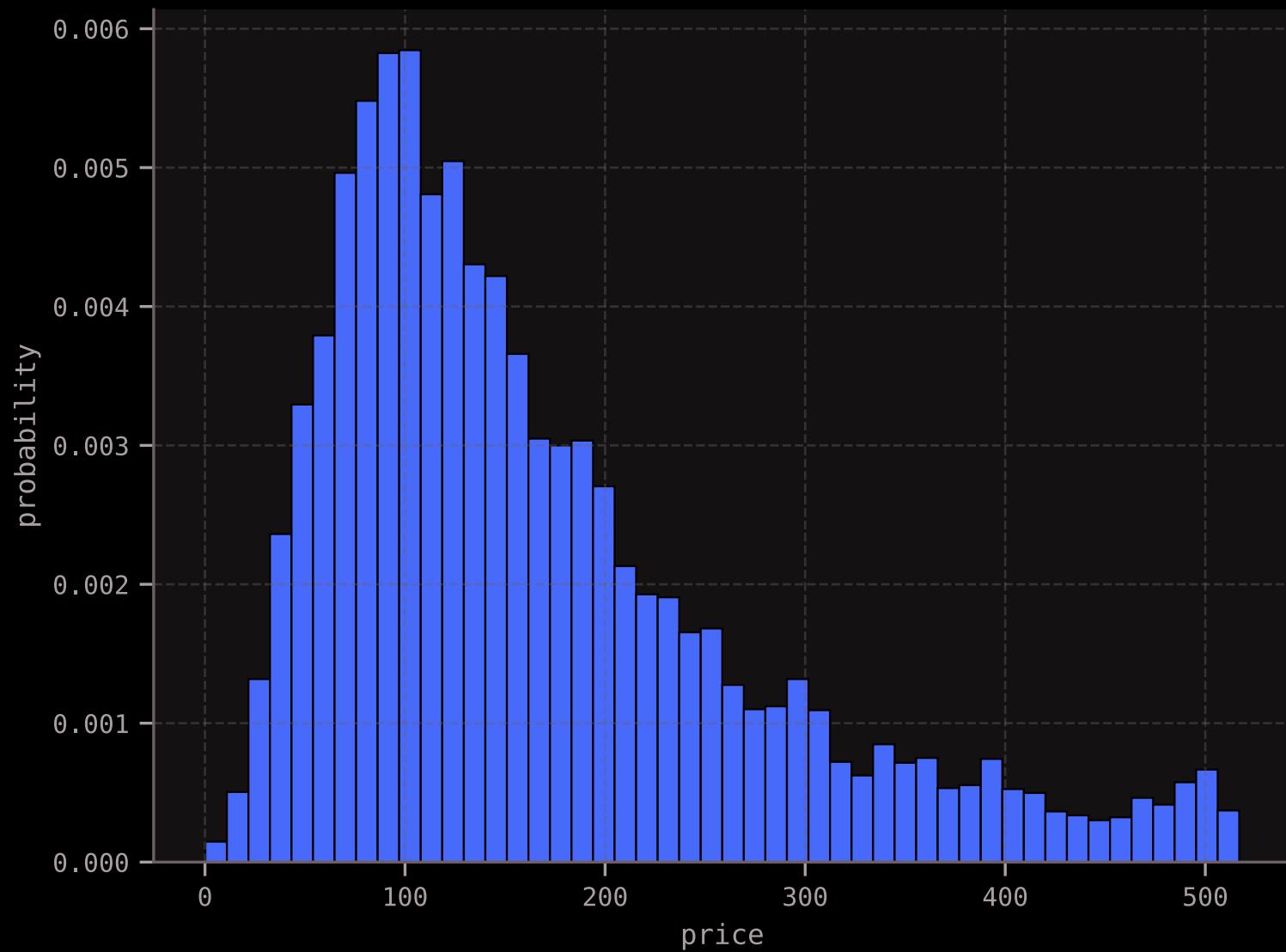


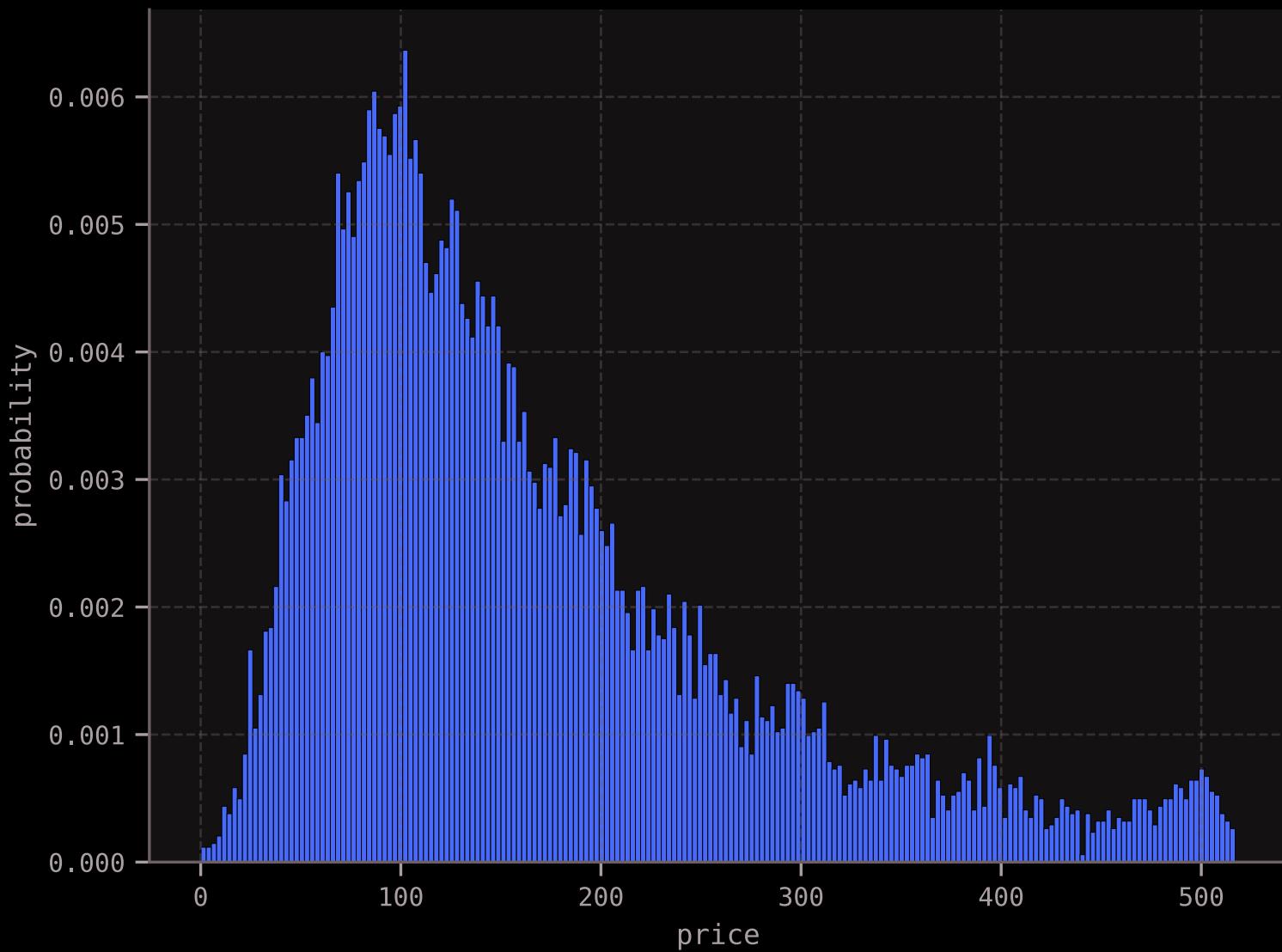
dataset in bins

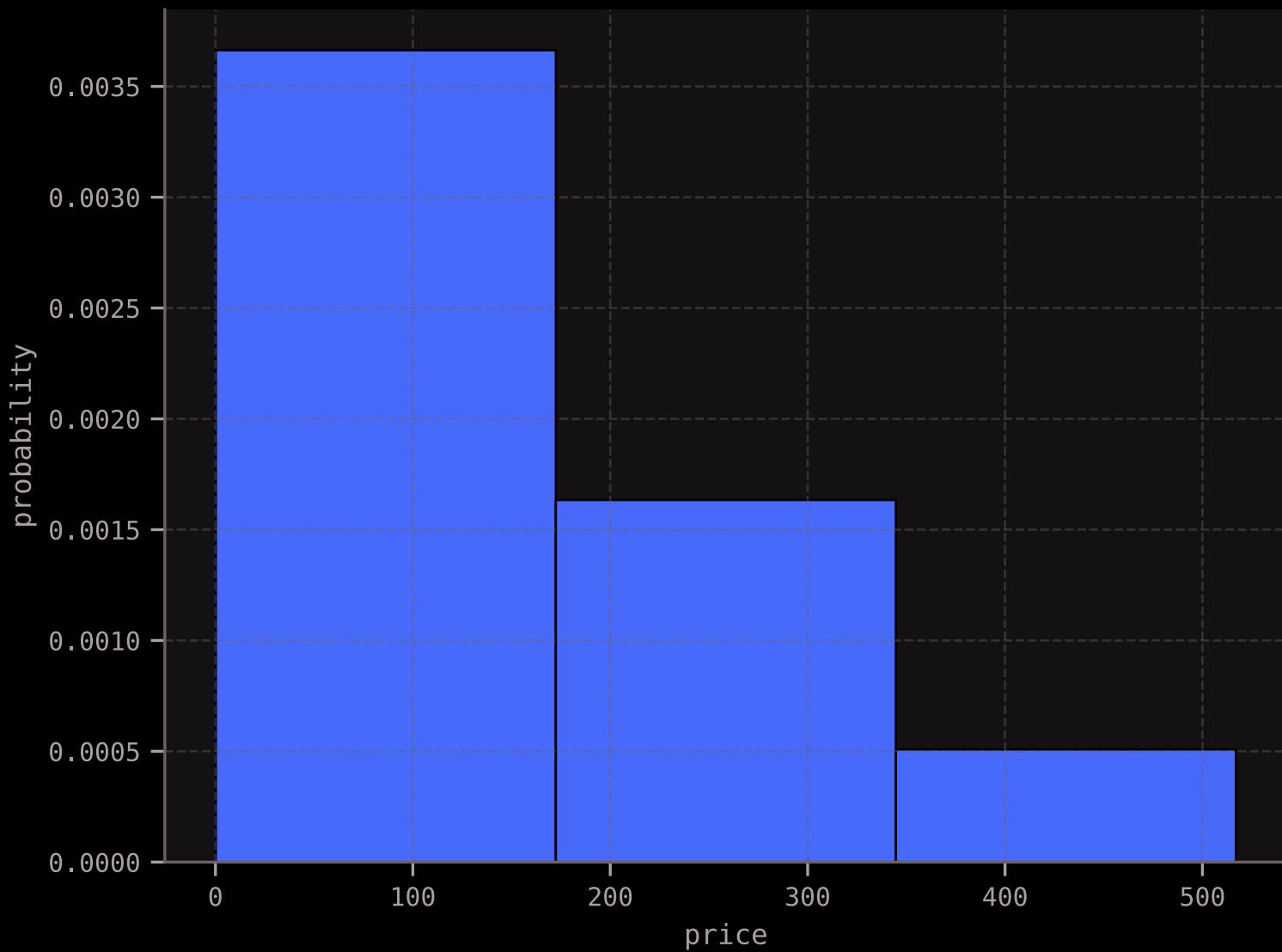


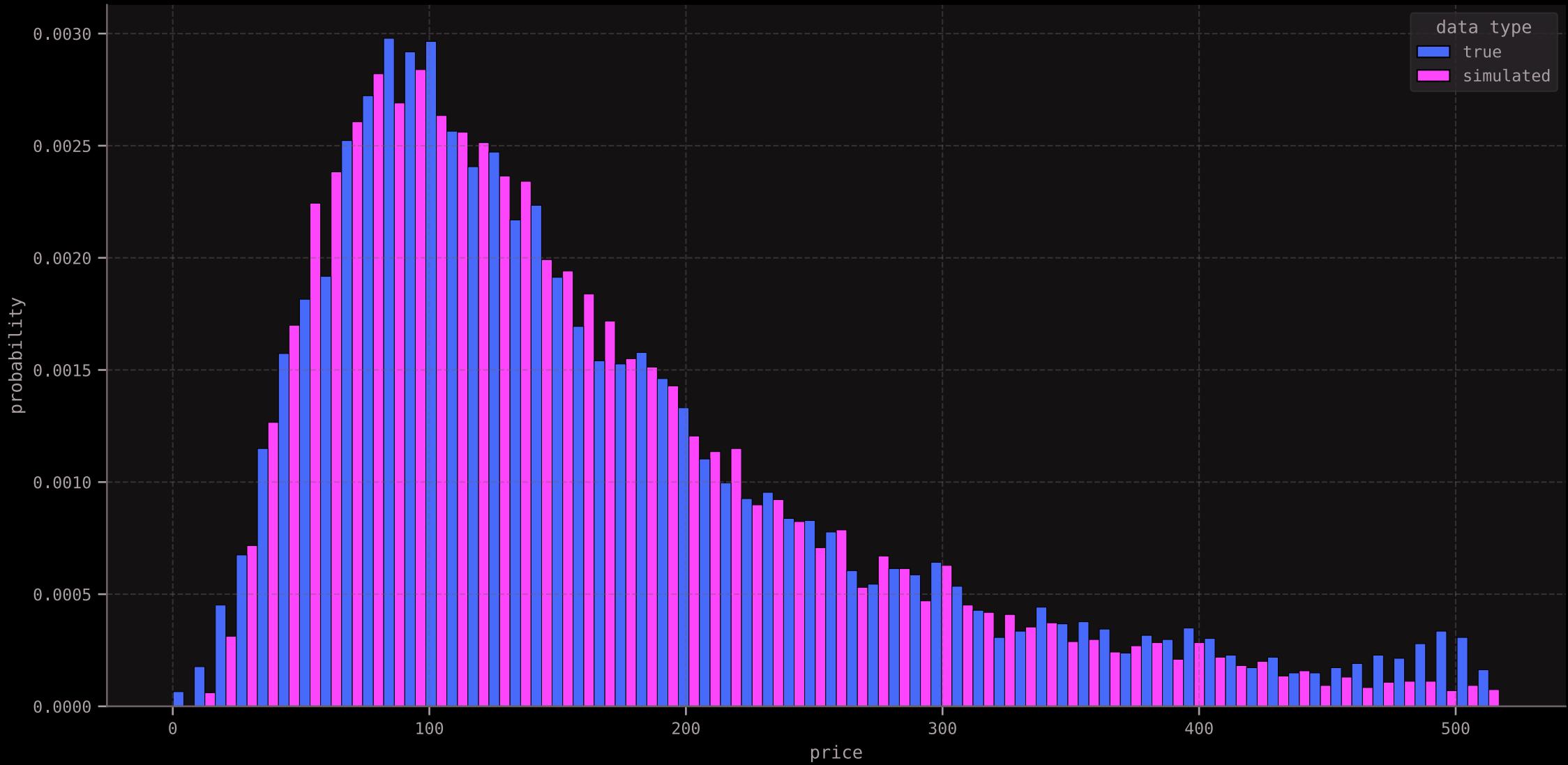
histogram





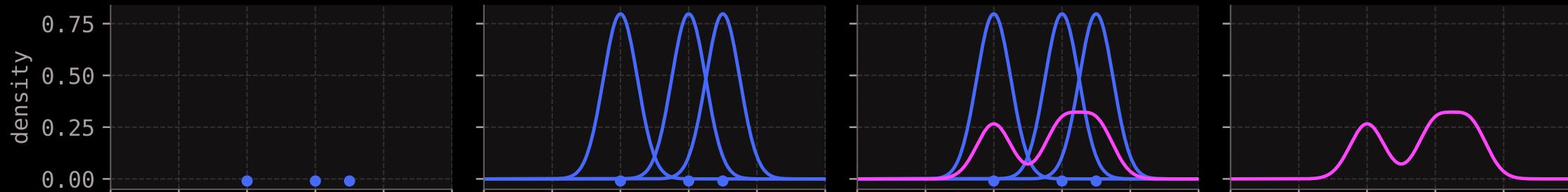




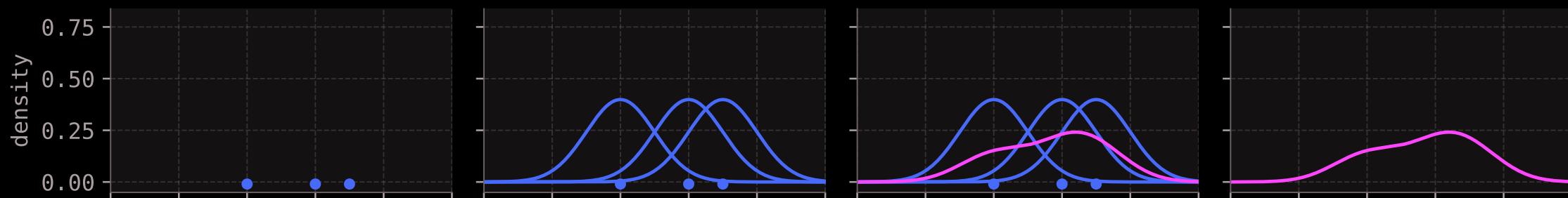


6.4. Kernel density estimation

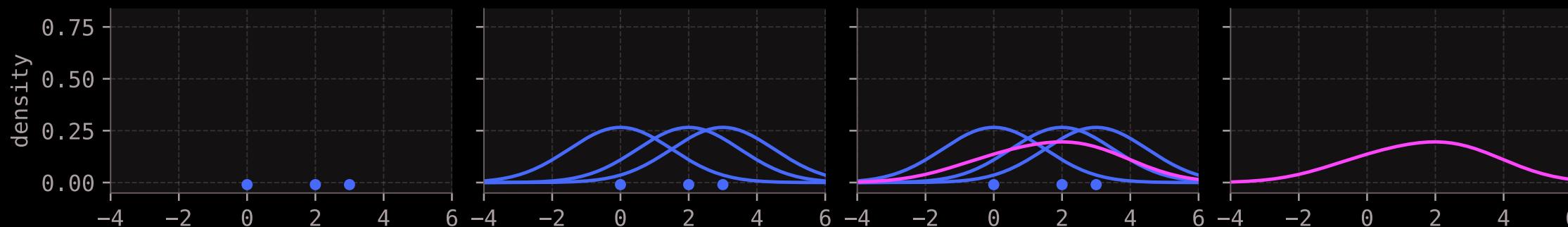
bandwidth $h = 0.5$

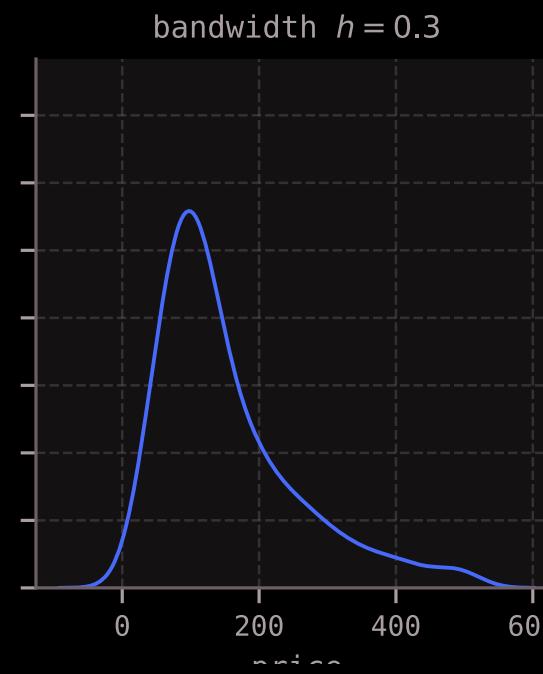
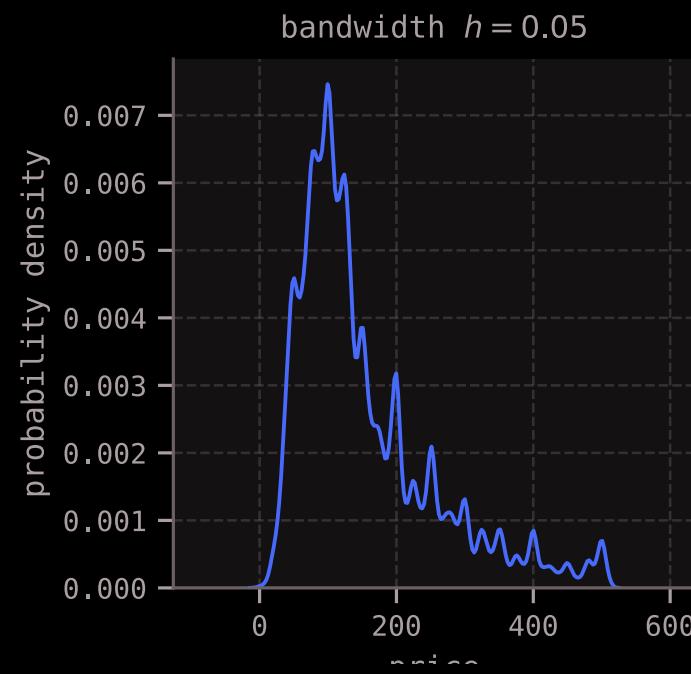


bandwidth $h = 1$



bandwidth $h = 1.5$





6.5. Empirical statistics



Definition 6.3

The *empirical mean* of a dataset x_1, x_2, \dots, x_m is defined to be the number

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i,$$

while the *empirical variance* is defined to be the number

$$s^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2.$$

The *empirical standard deviation* s is defined as the positive square root of the empirical variance, $s = \sqrt{s^2}$.

🔔 Definition 6.4

Let x_1, x_2, \dots, x_m be a dataset, written in non-decreasing order:

$$x_1 \leq x_2 \leq \dots \leq x_m. \quad (6.2)$$

For each $i = 1, 2, \dots, m$, the datapoint x_i is called the *empirical q_i -quantile* where

$$q_i = \frac{i - 1}{m - 1}. \quad (6.3)$$



Definition 6.5

The *empirical interquartile range (empirical IQR)* of a dataset x_1, x_2, \dots, x_m is the difference

(empirical 0.75-quantile) – (empirical 0.25-quantile).

Definition 6.6

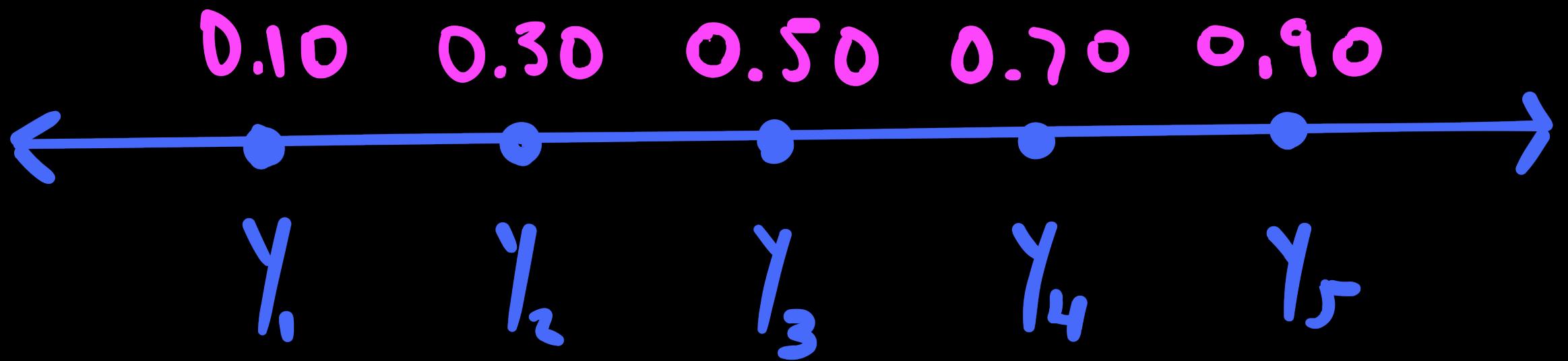
Let x_1, x_2, \dots, x_m be a dataset. Then a data point x_i is called an *outlier* if it is above an upper threshold value

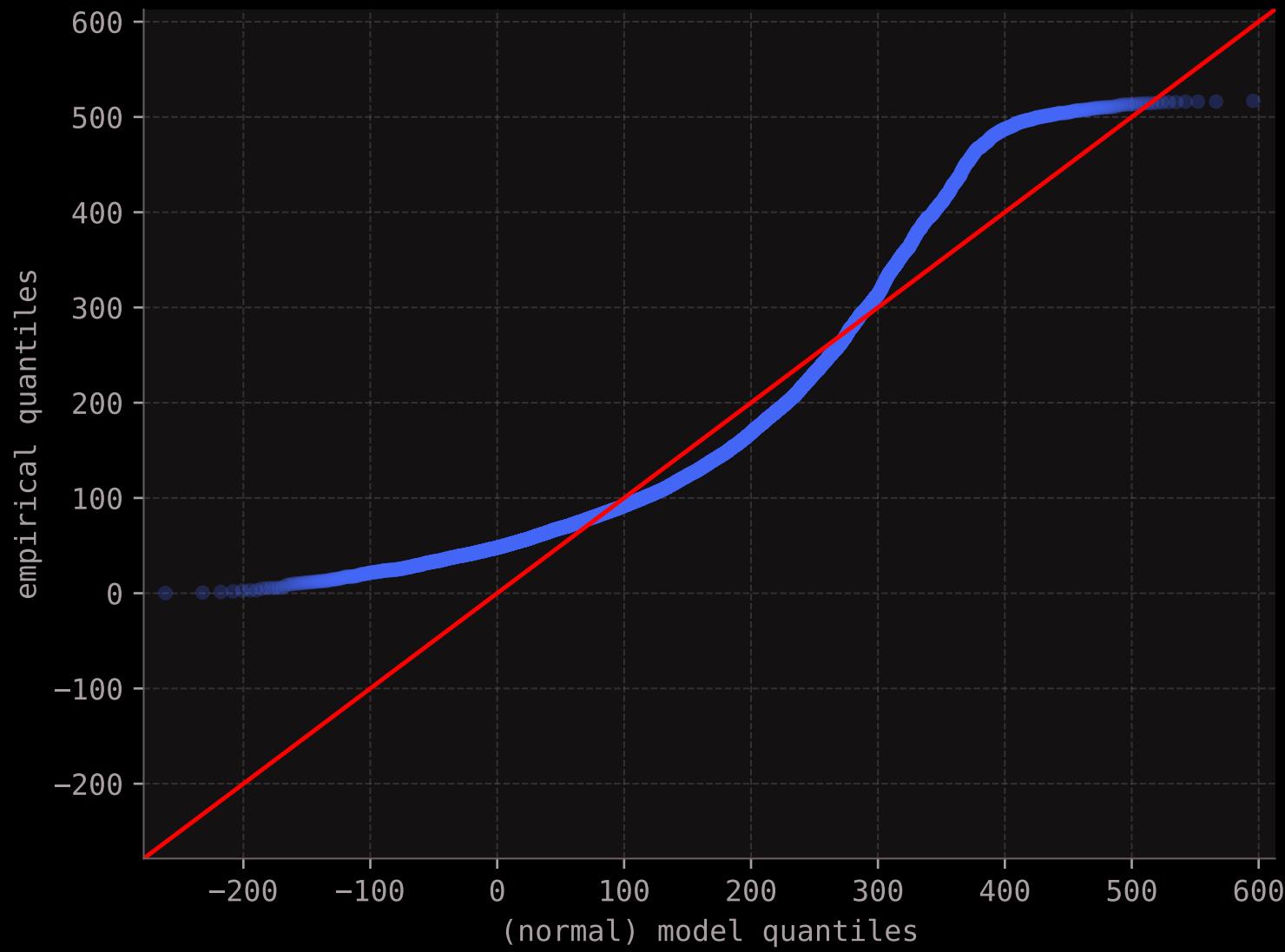
$$x_i > (\text{empirical 0.75-quantile}) + 1.5 \times (\text{empirical IQR}),$$

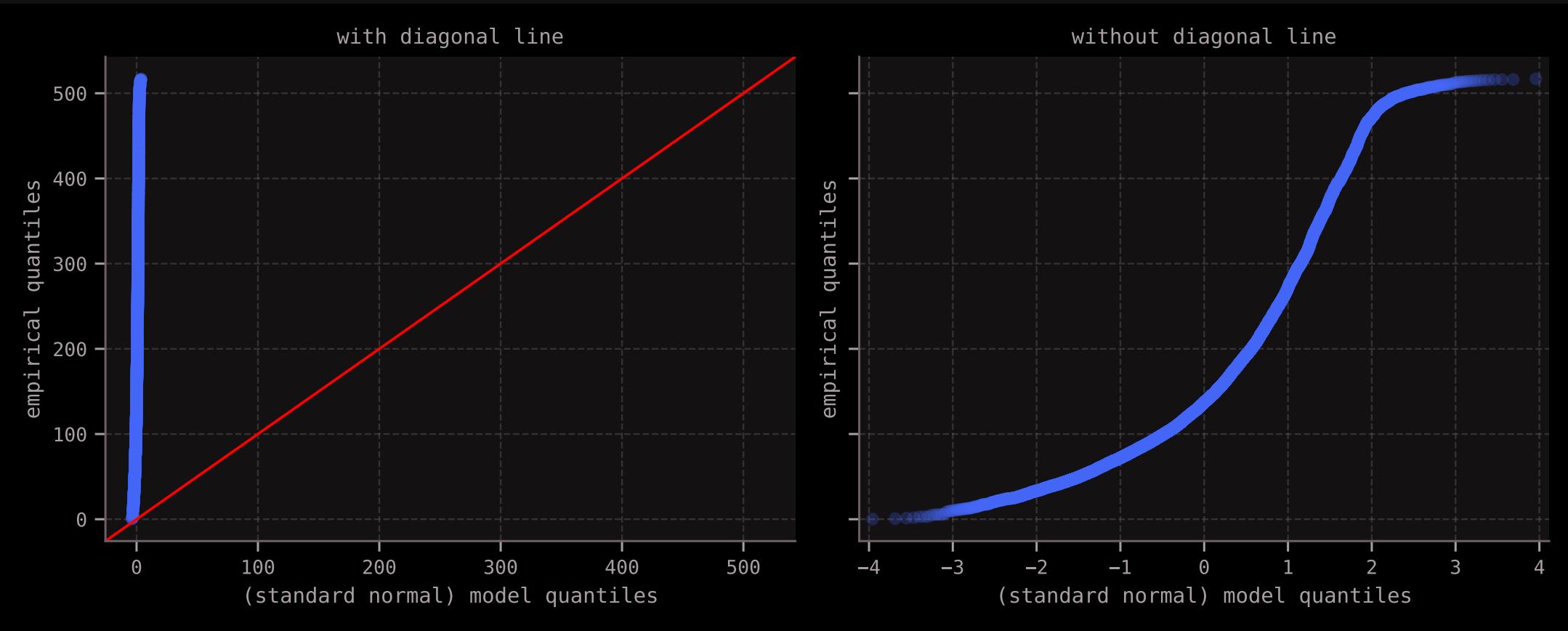
or if it is below a lower threshold value

$$x_i < (\text{empirical 0.25-quantile}) - 1.5 \times (\text{empirical IQR}).$$

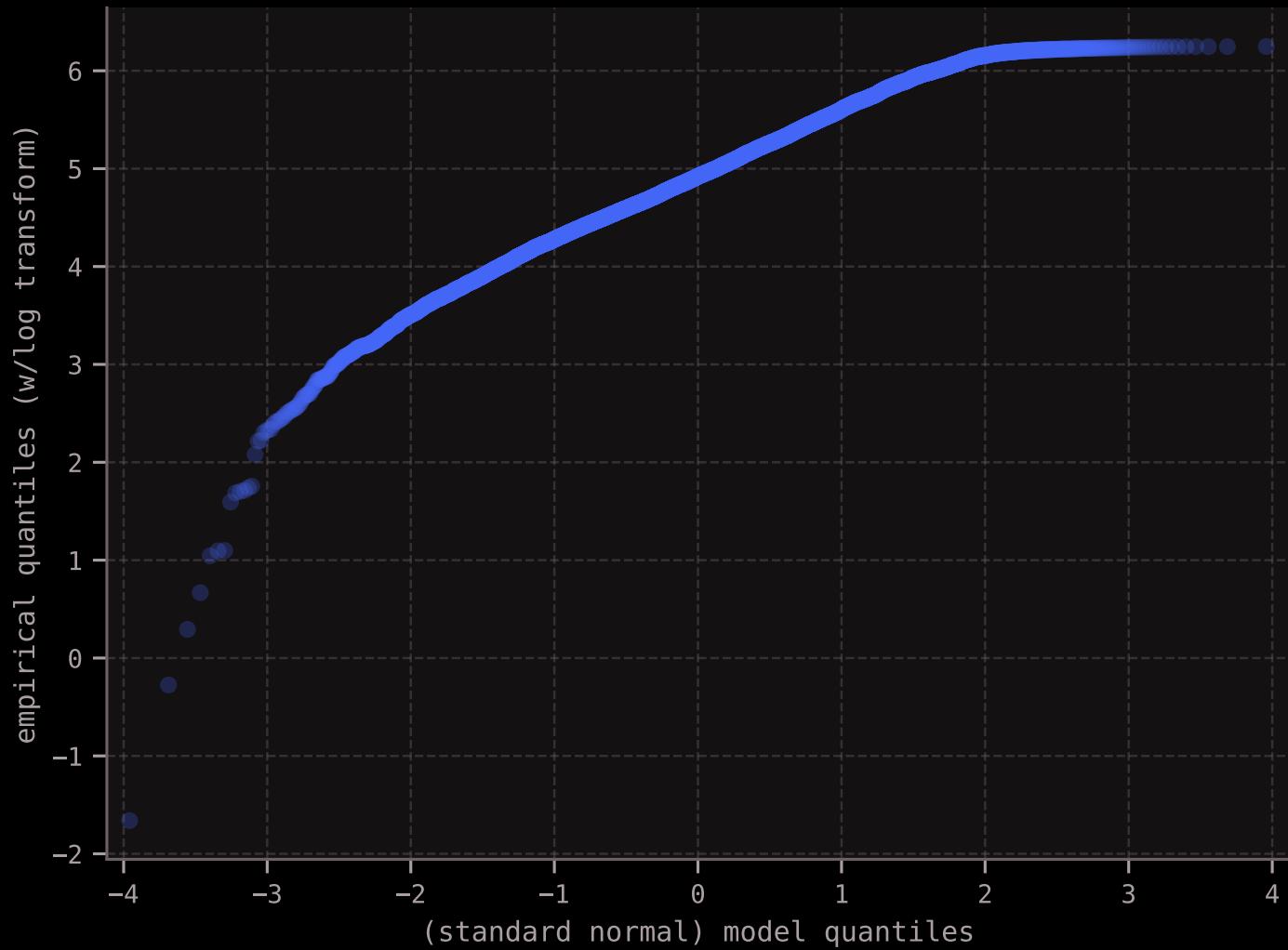
6.6. QQ-plots

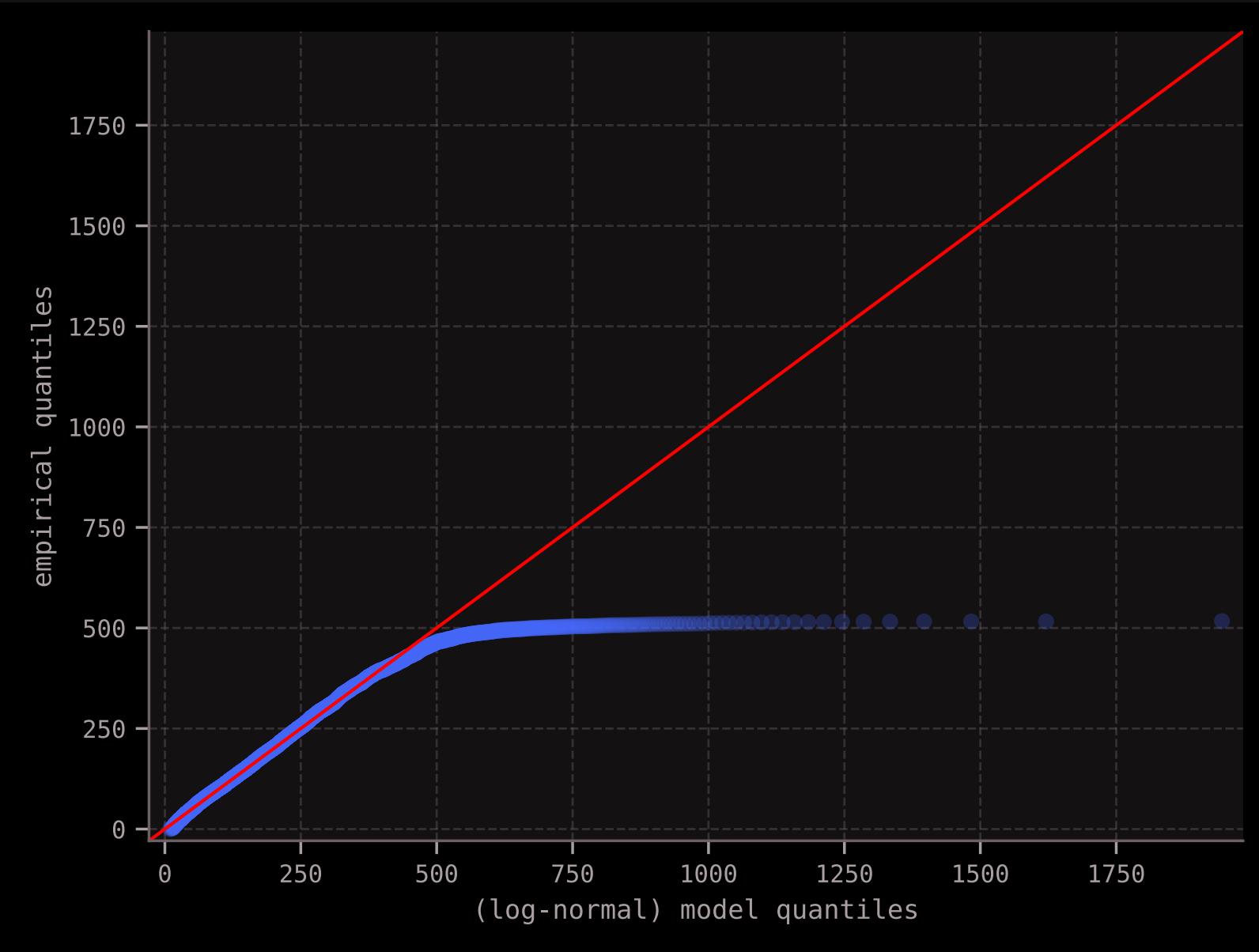


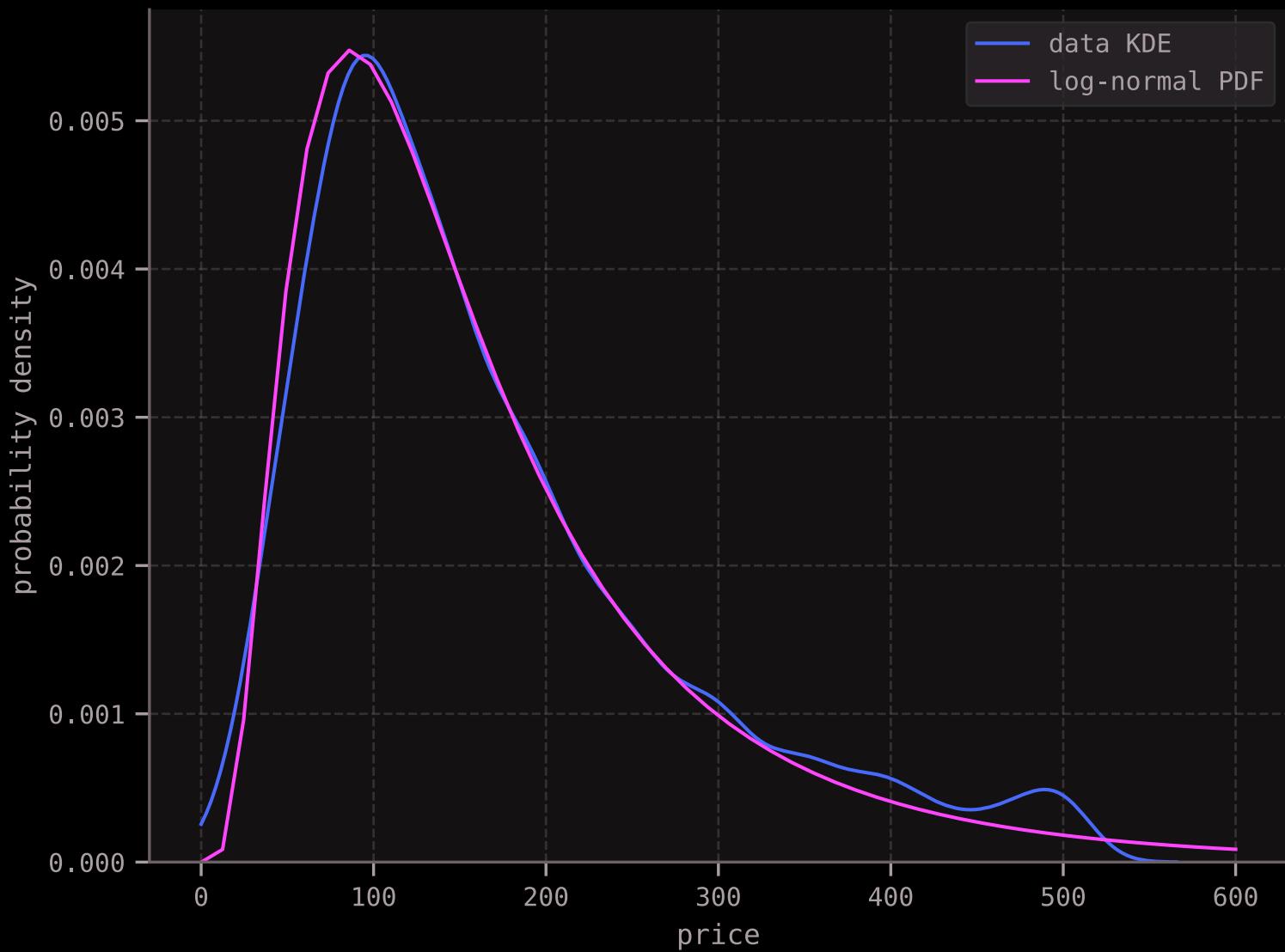




Observation: What we are looking for in the (standard normal) QQ-plot on the right is whether the scattered points fall along *some* straight line $y = ax + b$ (with $a > 0$). If they do, then the data is fit well by the normal distribution $\mathcal{N}(b, a^2)$.







6.7. Box plots and violin plots

