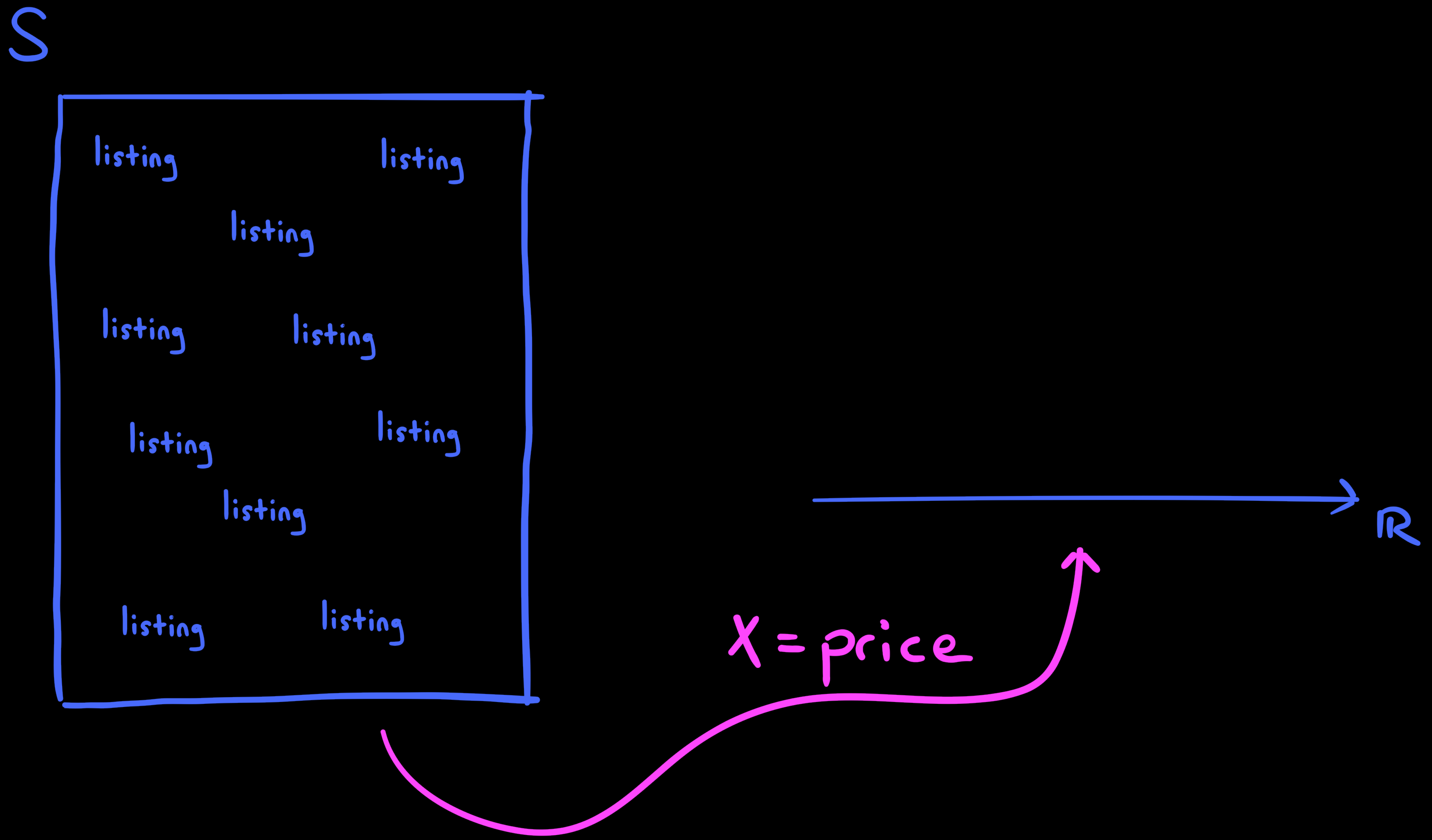
6. Connecting theory to practice: a first look at model building

model building

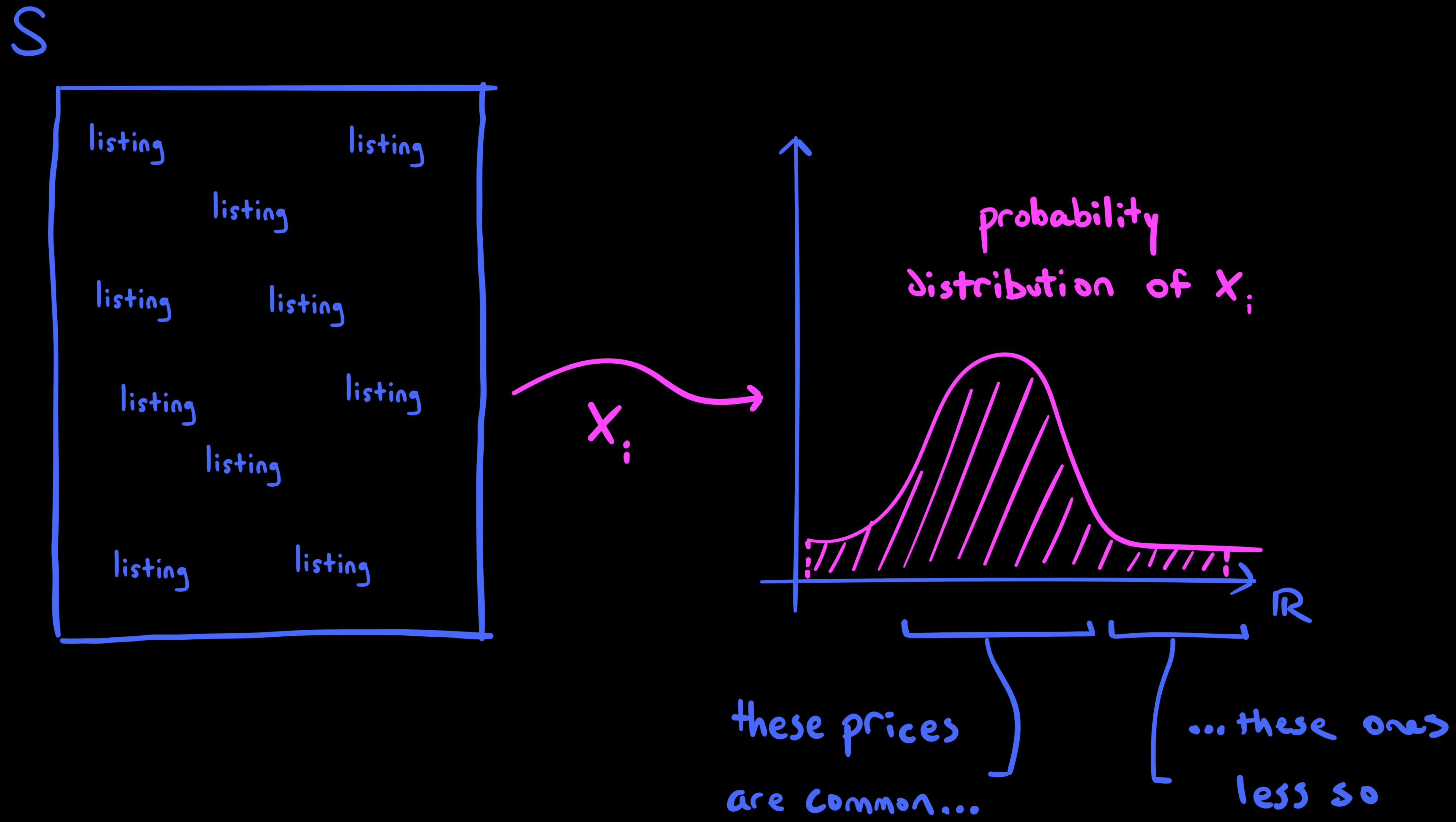
science

6.1. Data and random samples

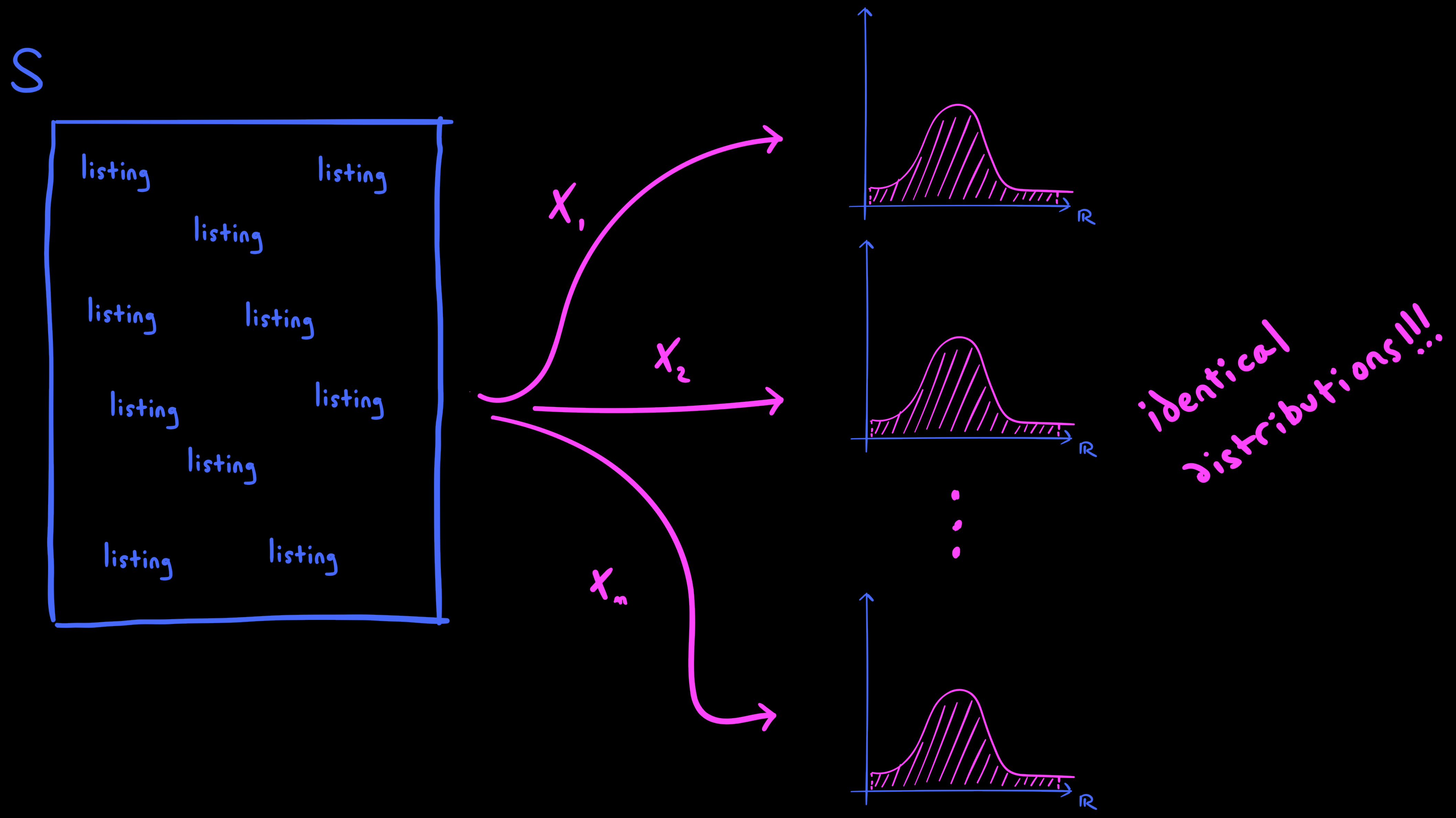


S
listing
listing
listing
listing
listing
listing
listing
listing
listing
listing
X = price
R

$$\begin{array}{ccccc} (\text{listing } 1, & \text{listing } 2, & \text{listing } 3, & \ldots, & \text{listing } m) \\ X_1 \Big\downarrow & X_2 \Big\downarrow & X_3 \Big\downarrow & & X_m \Big\downarrow \\ (\text{price } 1, & \text{price } 2, & \text{price } 3, & \ldots, & \text{price } m) \end{array}$$



S
listing
listing
listing
listing
listing
listing
listing
listing
listing
listing
X_i
probability distribution of X_i
R
these prices are common...
... these ones less so


S
listing
listing
listing
listing
listing
listing
listing
listing
listing
listing
X_1
X_2
X_m
$\mathbb{R}$
$\mathbb{R}$
$\mathbb{R}$
identical distributions!!!

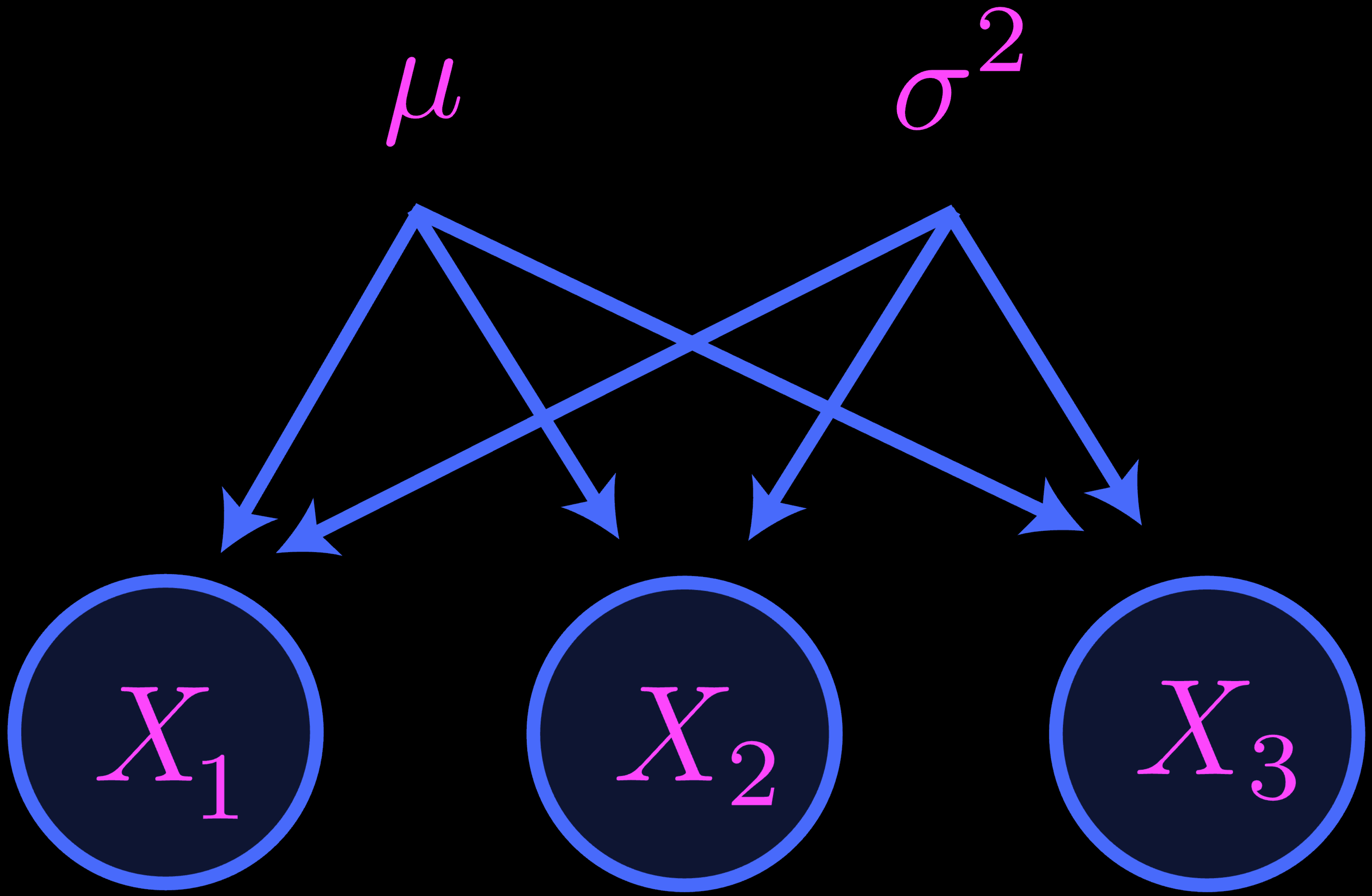
Definition 6.1

Let $X_1, X_2, \ldots, X_m$ 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 $x_1, x_2, \ldots, x_m$ 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



μ
σ^2
X_1
X_2
X_3

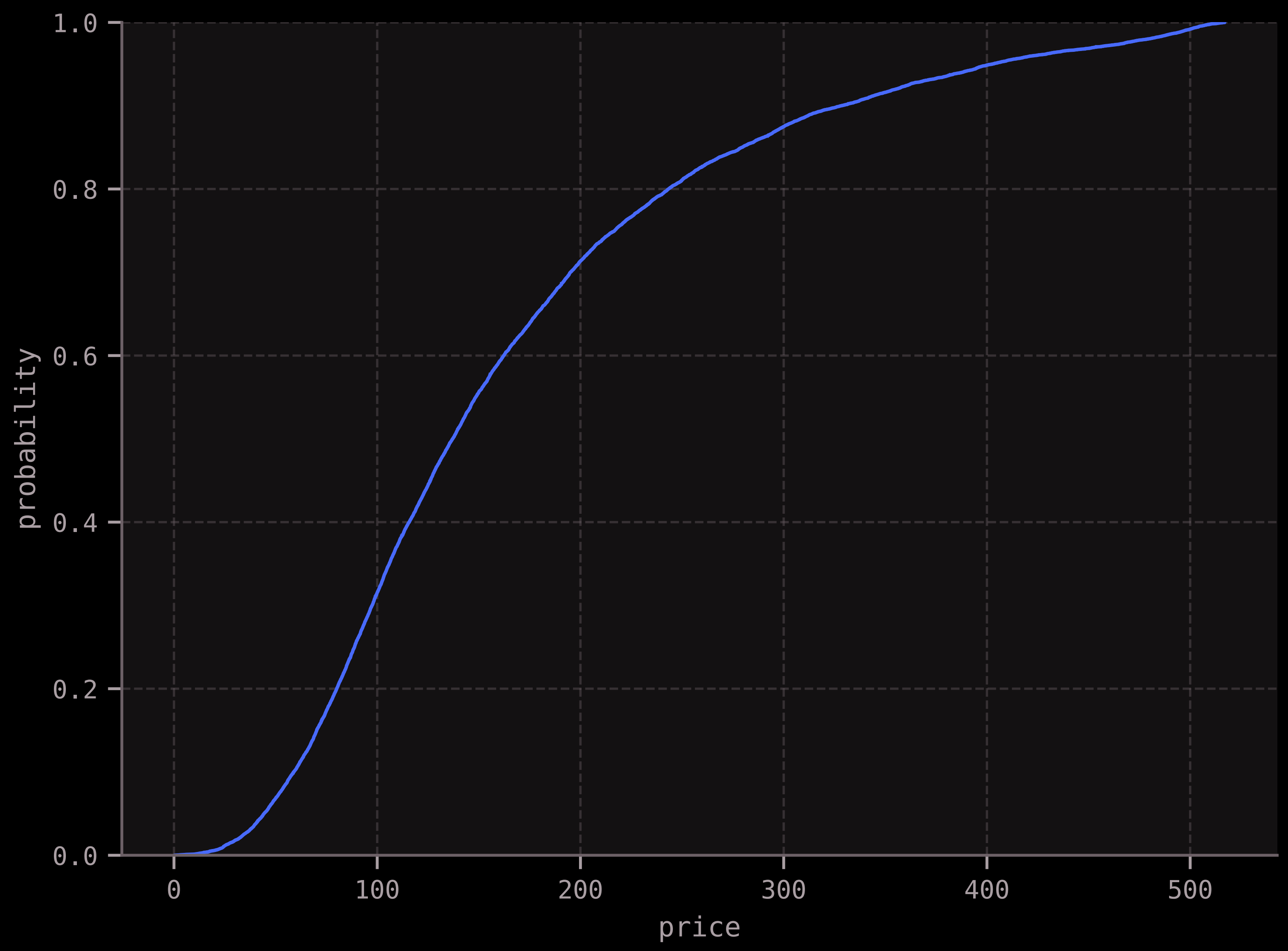
Definition 6.2

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

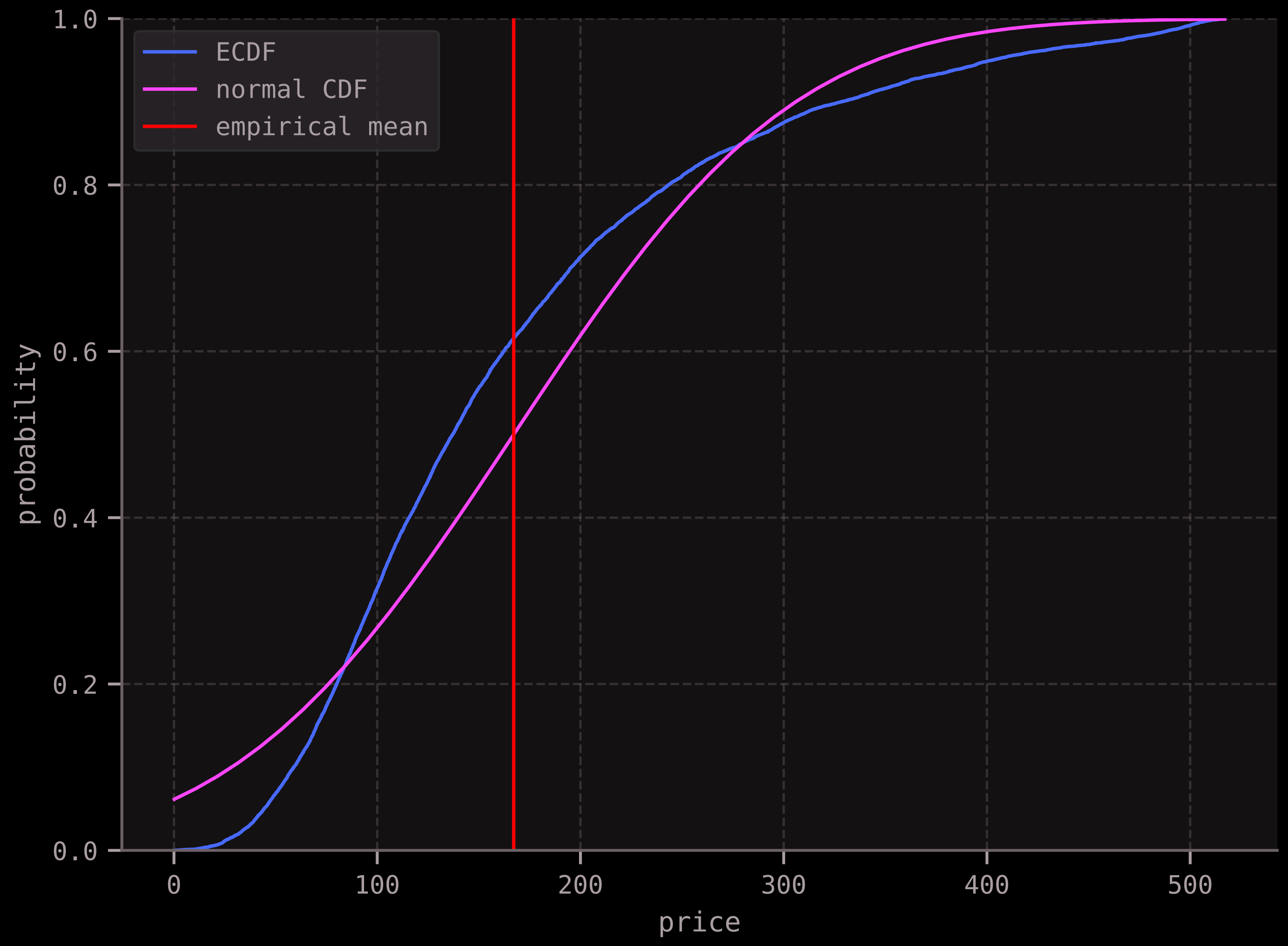
$$p(x) = \frac{\text{number of data points } x_i \text{ that match } x}{m}.$$

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

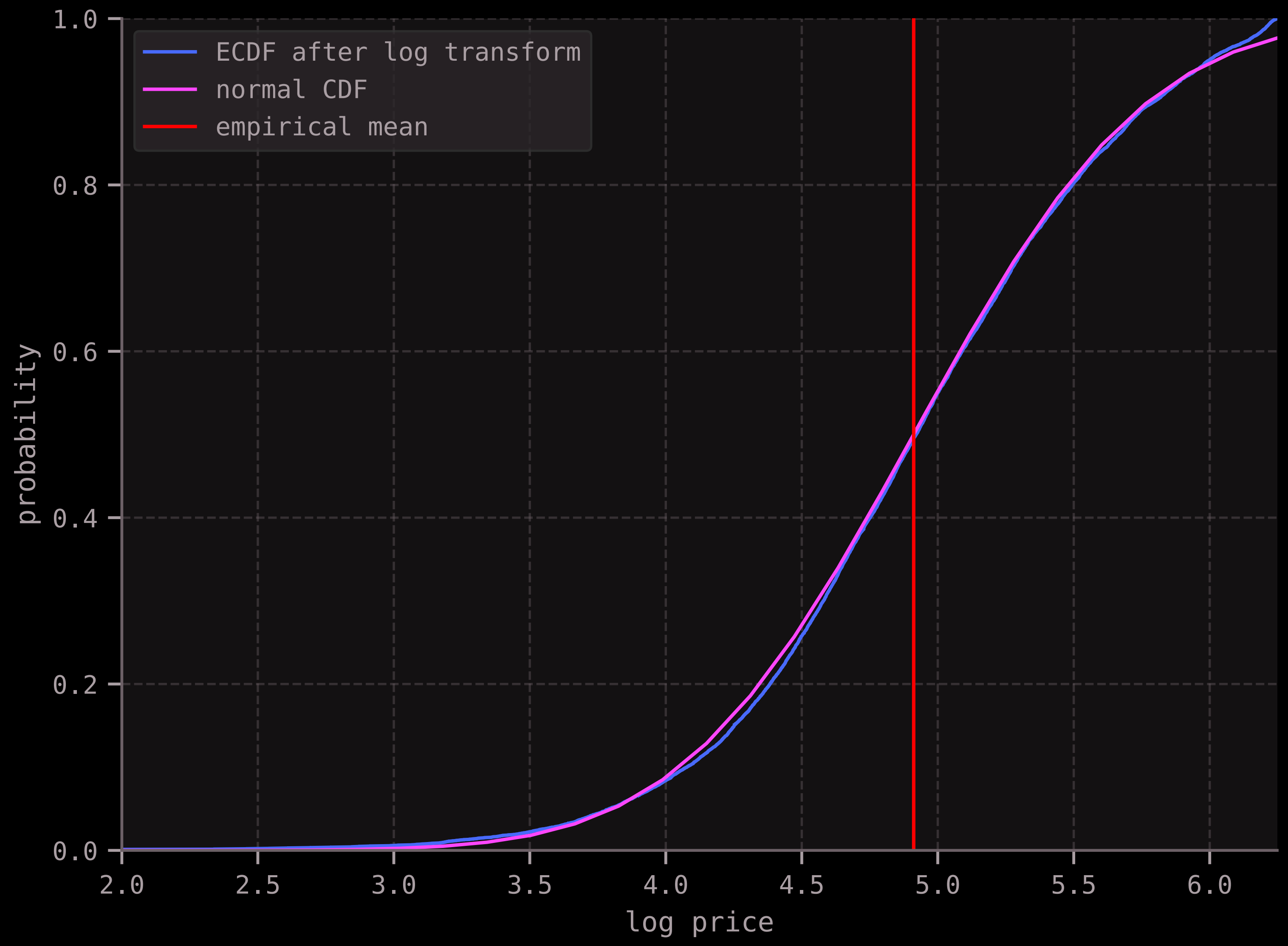
$$F(x) = \sum_{x^\star \leq x} p(x^\star) = \frac{\text{number of data points } x_i \text{ with } x_i \leq x}{m}.$$



1.0
0.8
0.6
0.4
0.2
0.0
probability
0
100
200
300
400
500
price

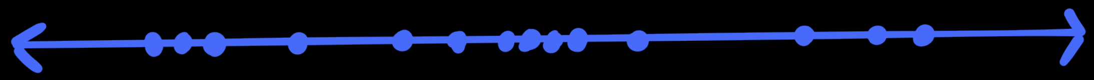


ECDF
normal CDF
empirical mean
probability
price
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

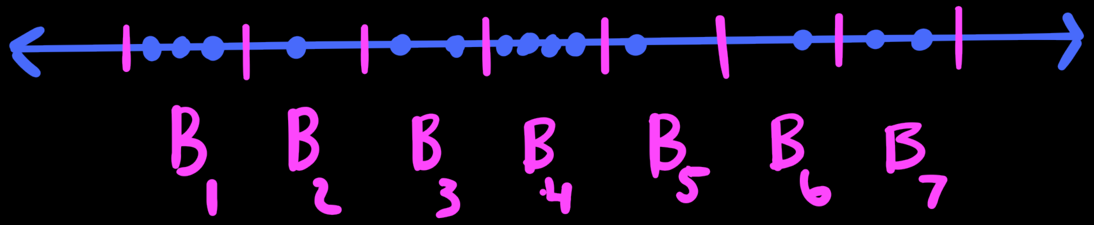

ECDF after log transform
normal CDF
empirical mean
1.0
0.8
0.6
0.4
0.2
0.0
probability
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
log price

6.3. Histograms

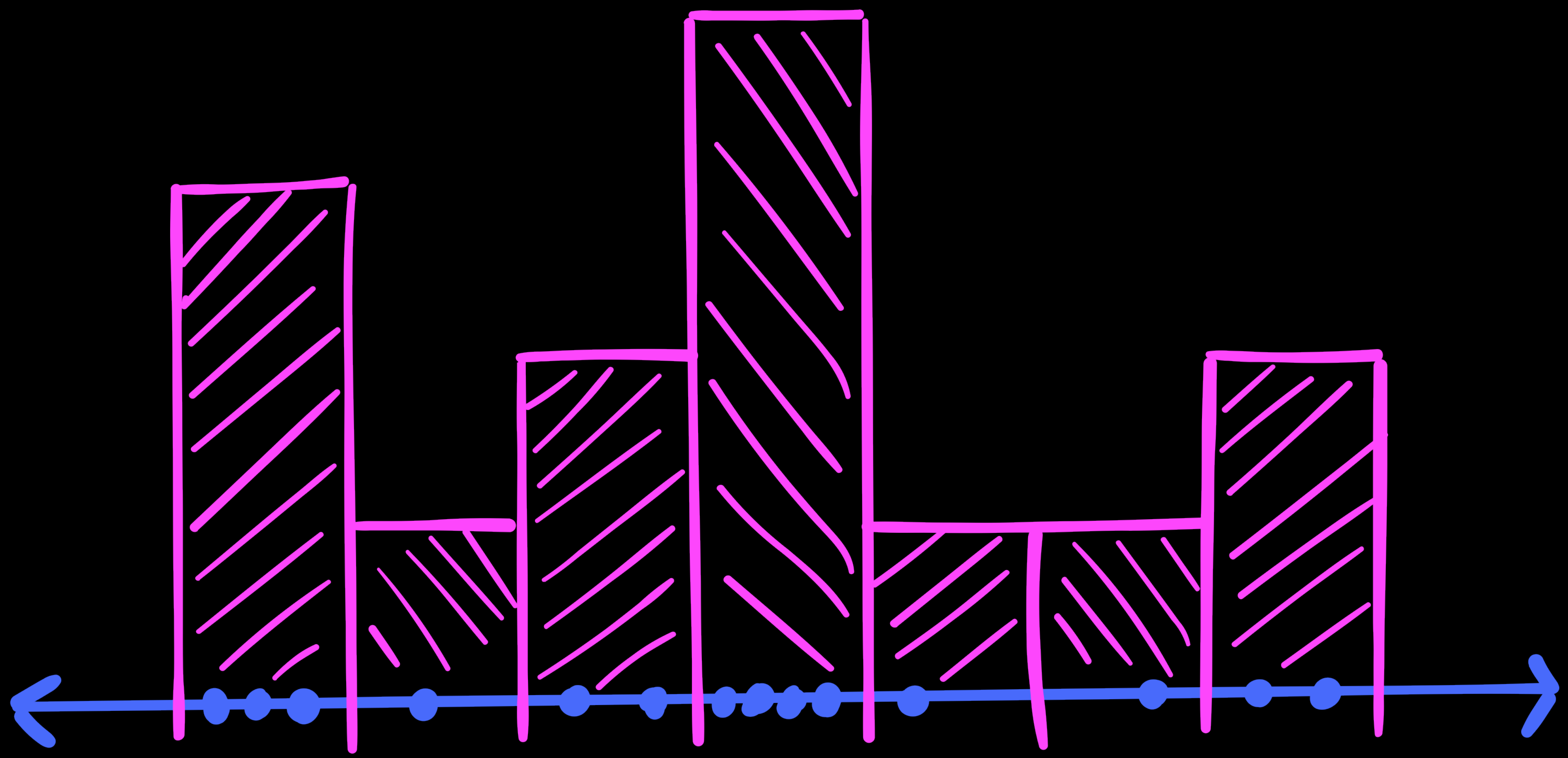
dataset

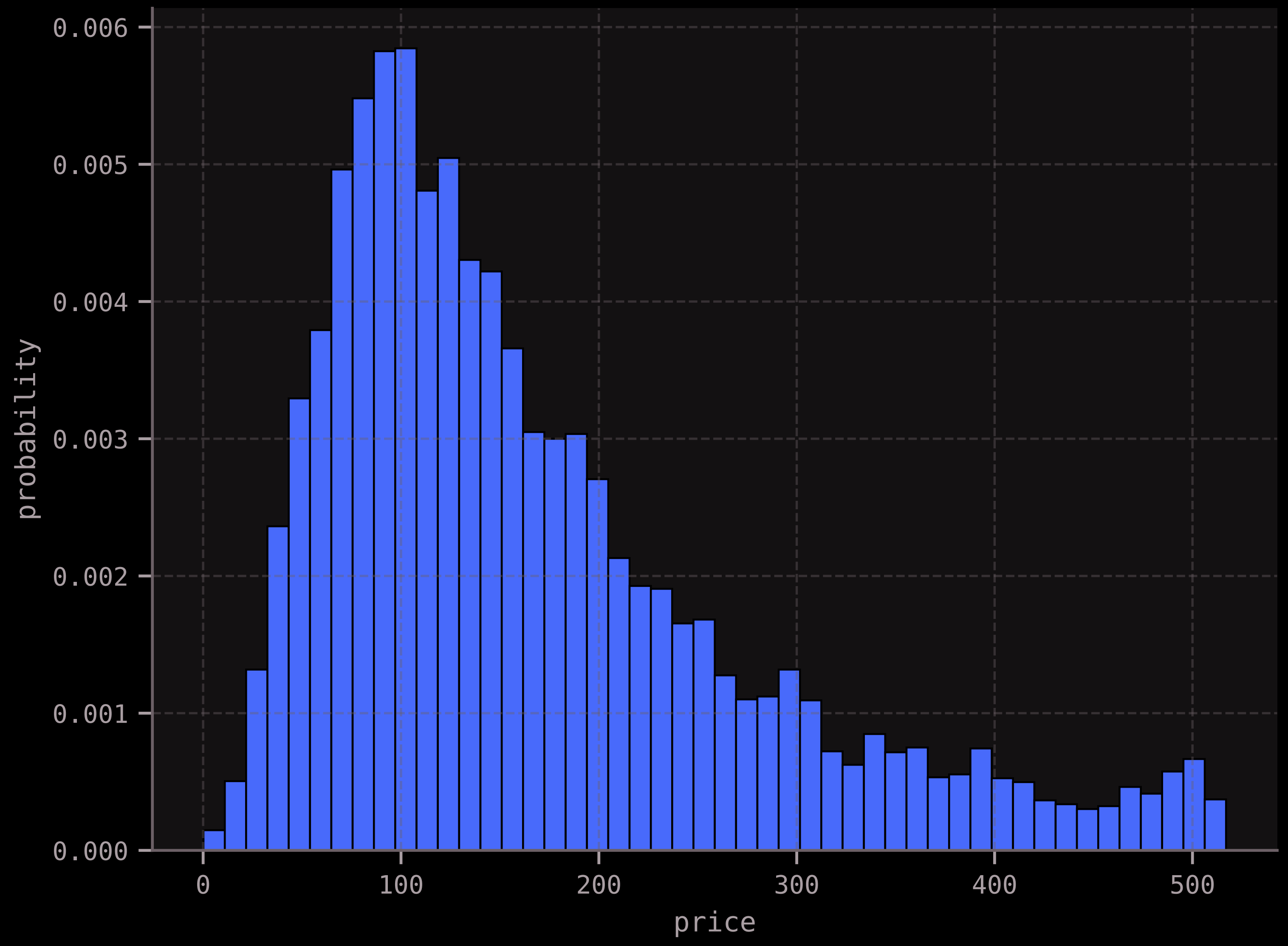


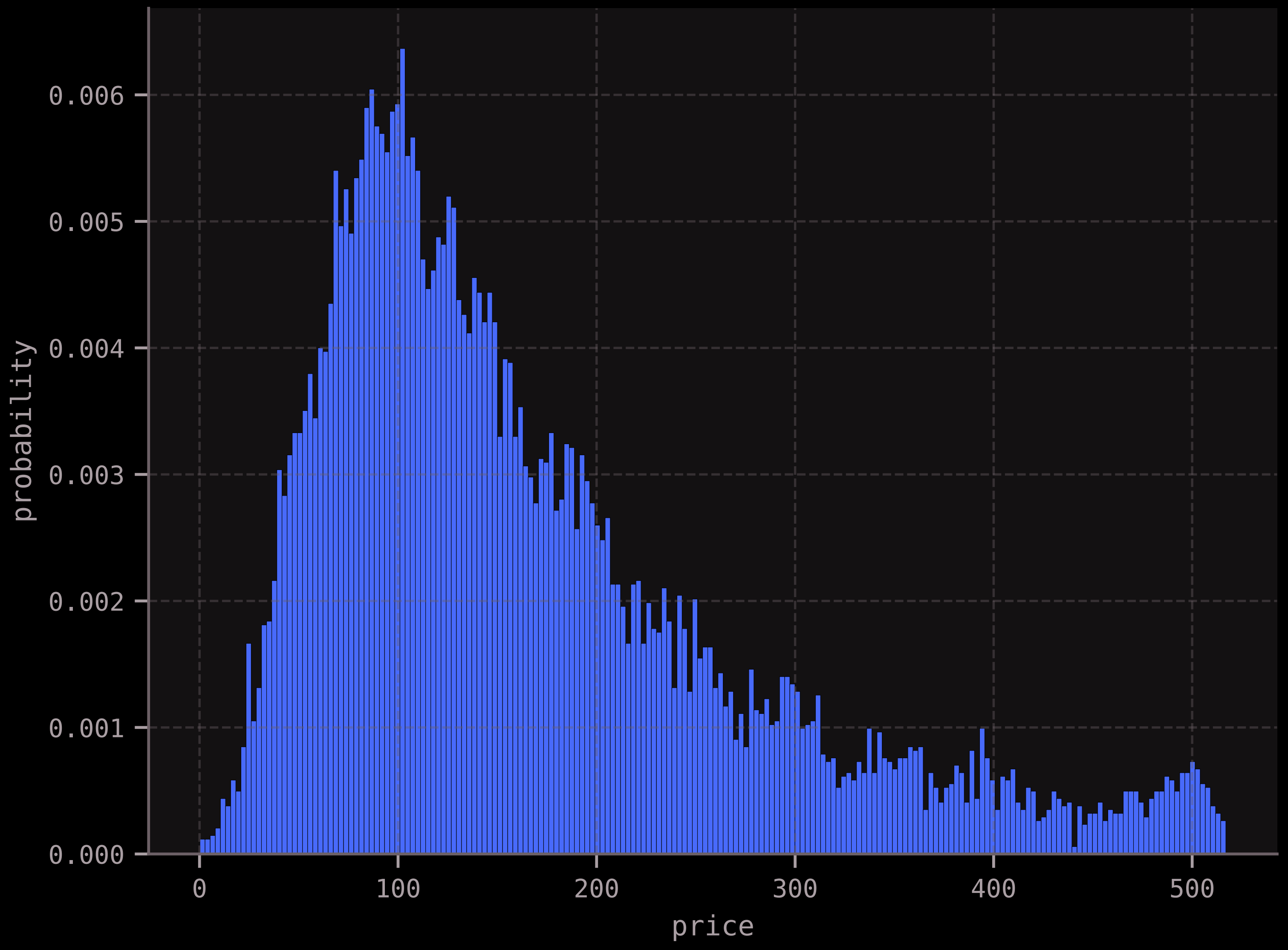
dataset in bins



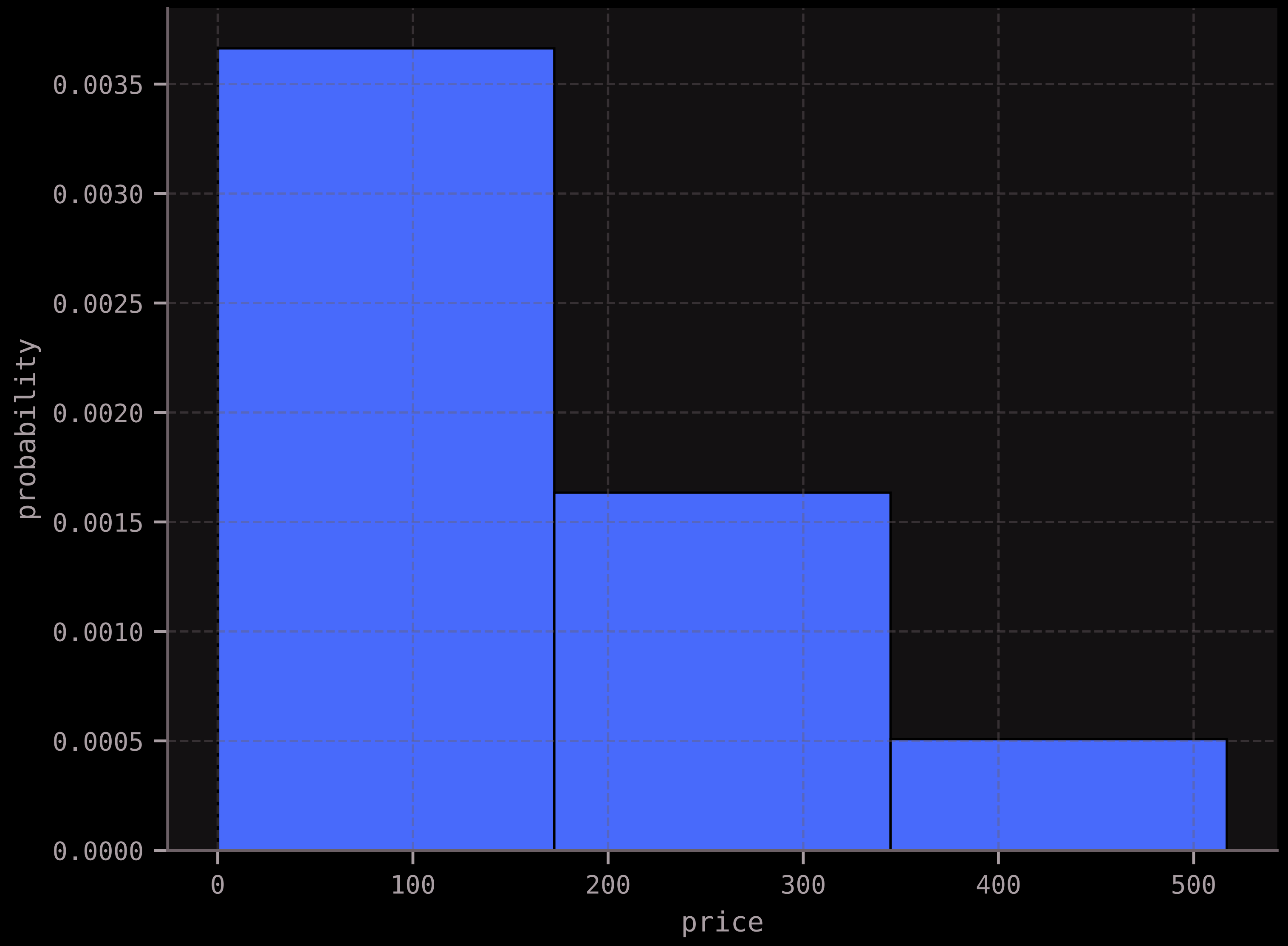
histogram



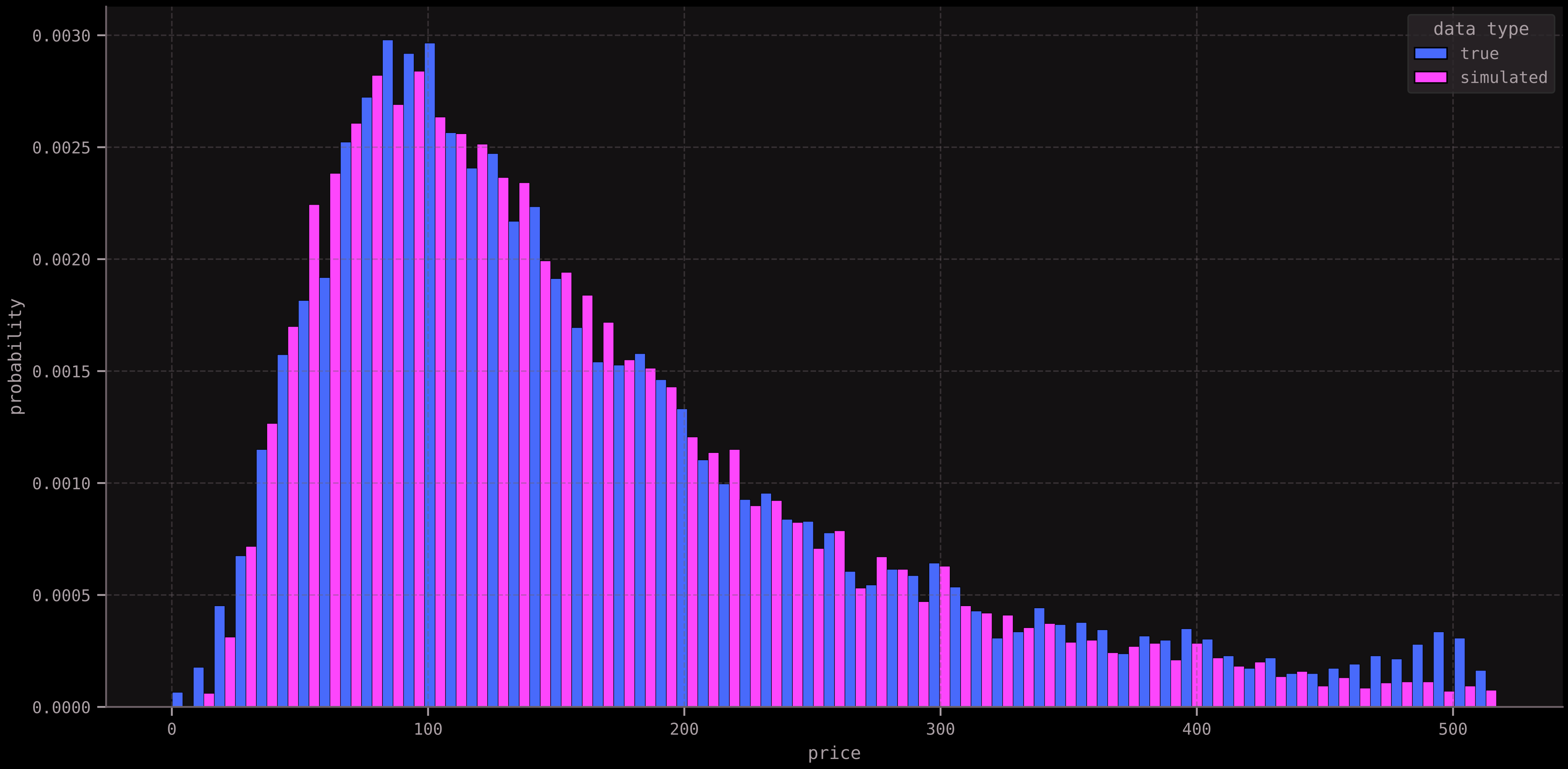

probability
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
100
200
300
400
500
price



probability
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400
500
price

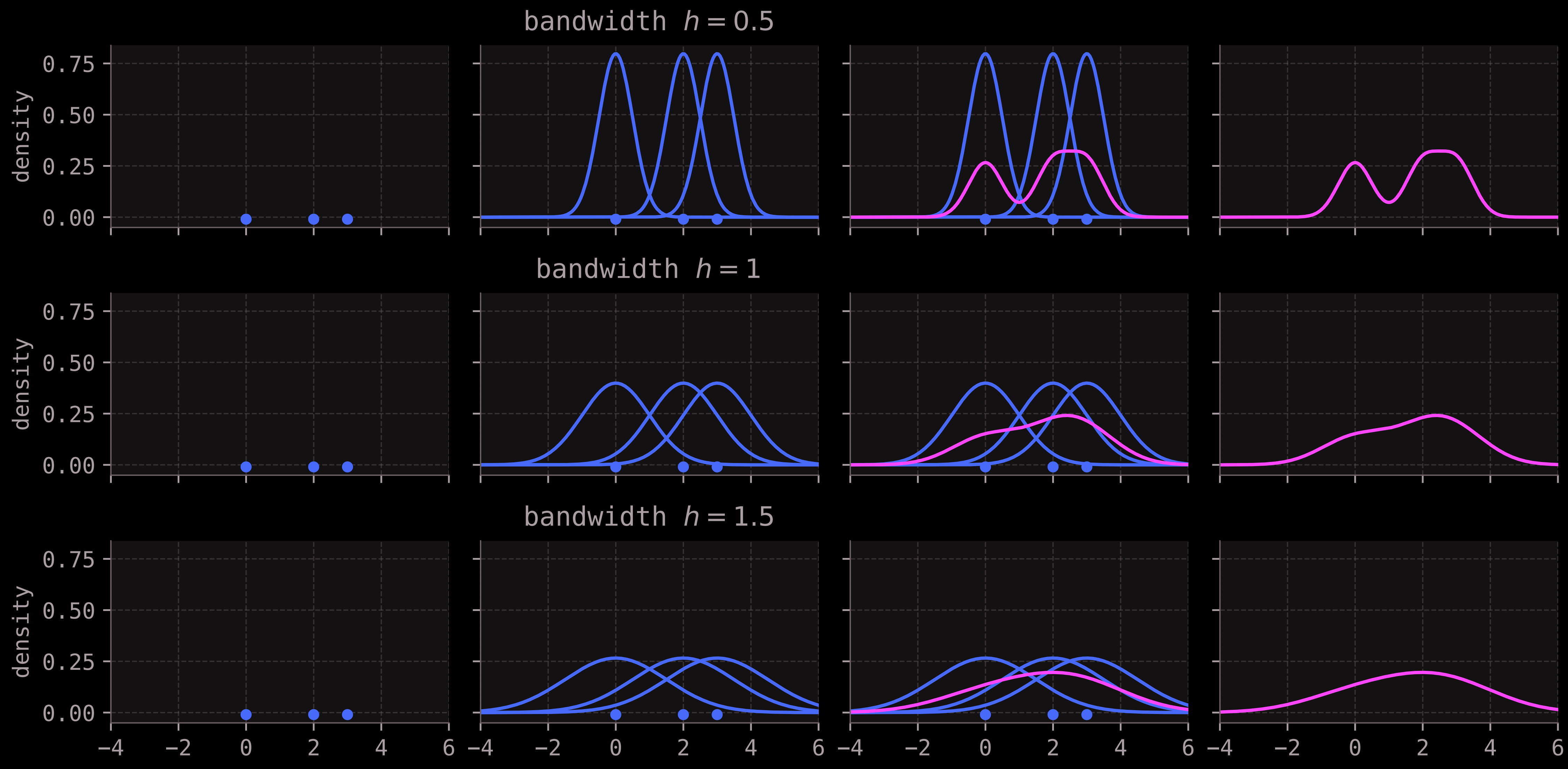


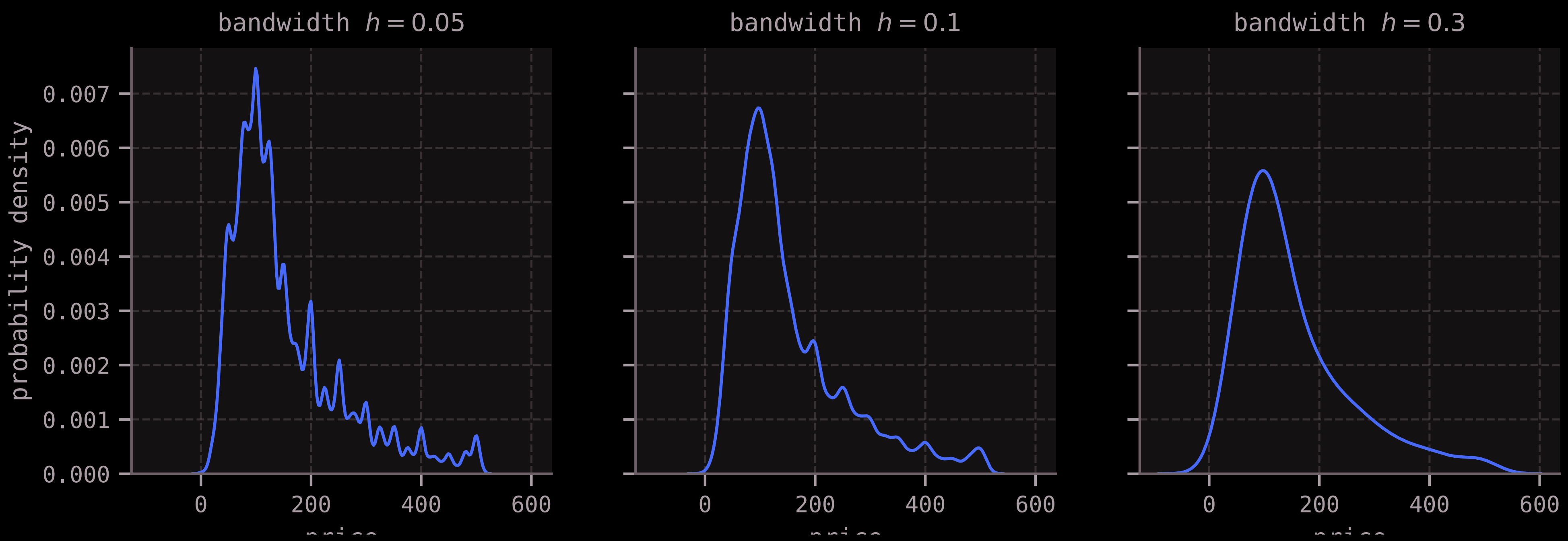
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
probability
0
100
200
300
400
500
price



data type
true
simulated
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
probability
0
100
200
300
400
500
price

6.4. Kernel density estimation


bandwidth $h = 0.5$
bandwidth $h = 1$
bandwidth $h = 1.5$
density
0.00
0.25
0.50
0.75
−4
−2
0
2
4
6


bandwidth $h = 0.05$
bandwidth $h = 0.1$
bandwidth $h = 0.3$
probability density
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
price

6.5. Empirical statistics

Definition 6.3

The *empirical mean* of a dataset $x_1, x_2, \ldots, x_m$ 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 $x_1, x_2, \ldots, x_m$ be a dataset, written in non-decreasing order:

$$x_1 \leq x_2 \leq \cdots \leq x_m. \tag{6.2}$$

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

$$q_i = \frac{i-1}{m-1}. \tag{6.3}$$

Definition 6.5

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

$$(\text{empirical } 0.75\text{-quantile}) - (\text{empirical } 0.25\text{-quantile}).$$

Definition 6.6

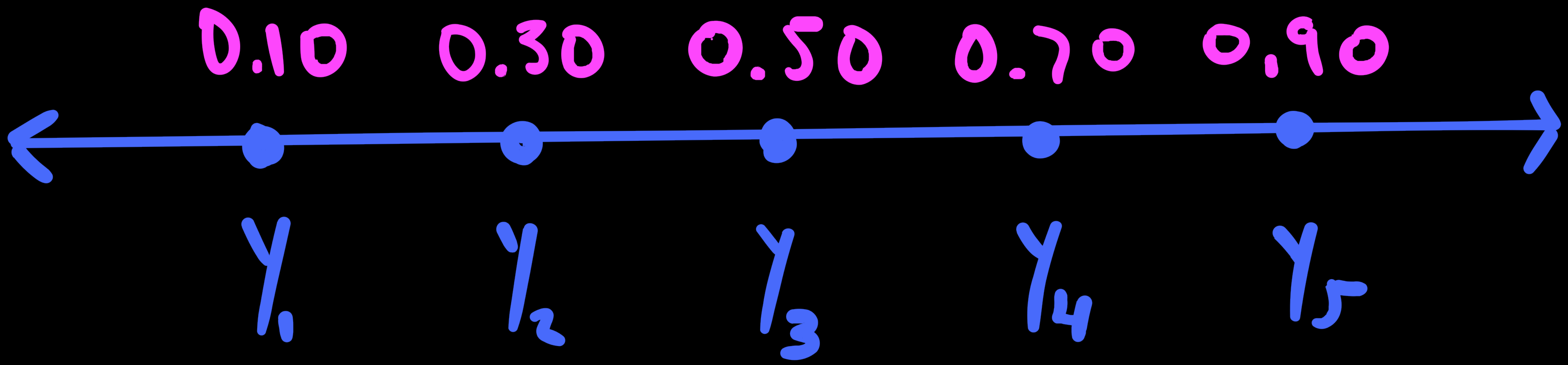
Let $x_1, x_2, \ldots, x_m$ be a dataset. Then a data point x_i is called an *outlier* if it is above an upper threshold value

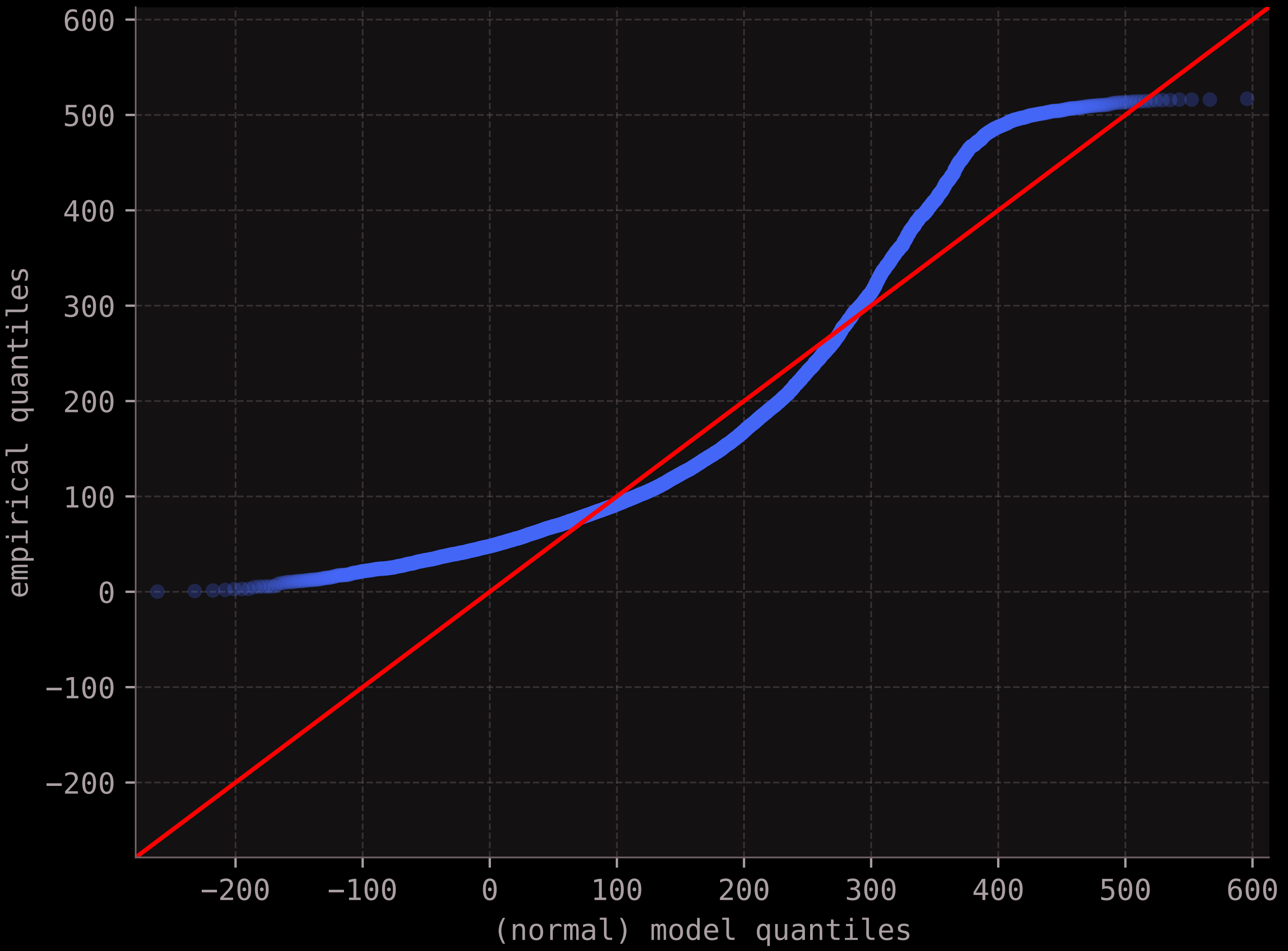
$$x_i > (\text{empirical } 0.75\text{-quantile}) + 1.5 \times (\text{empirical IQR}),$$

or if it is below a lower threshold value

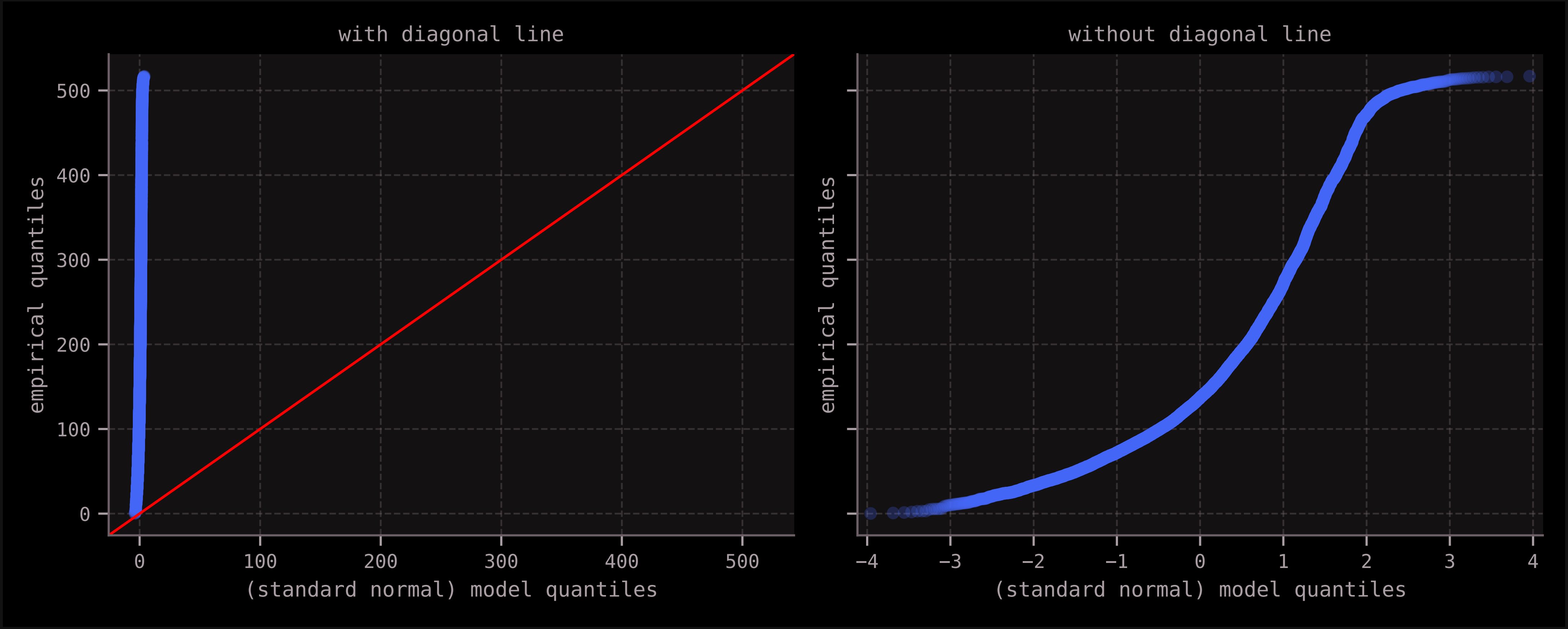
$$x_i < (\text{empirical } 0.25\text{-quantile}) - 1.5 \times (\text{empirical IQR}).$$

6.6. QQ-plots


0.10
0.30
0.50
0.70
0.90
y_1
y_2
y_3
y_4
y_5

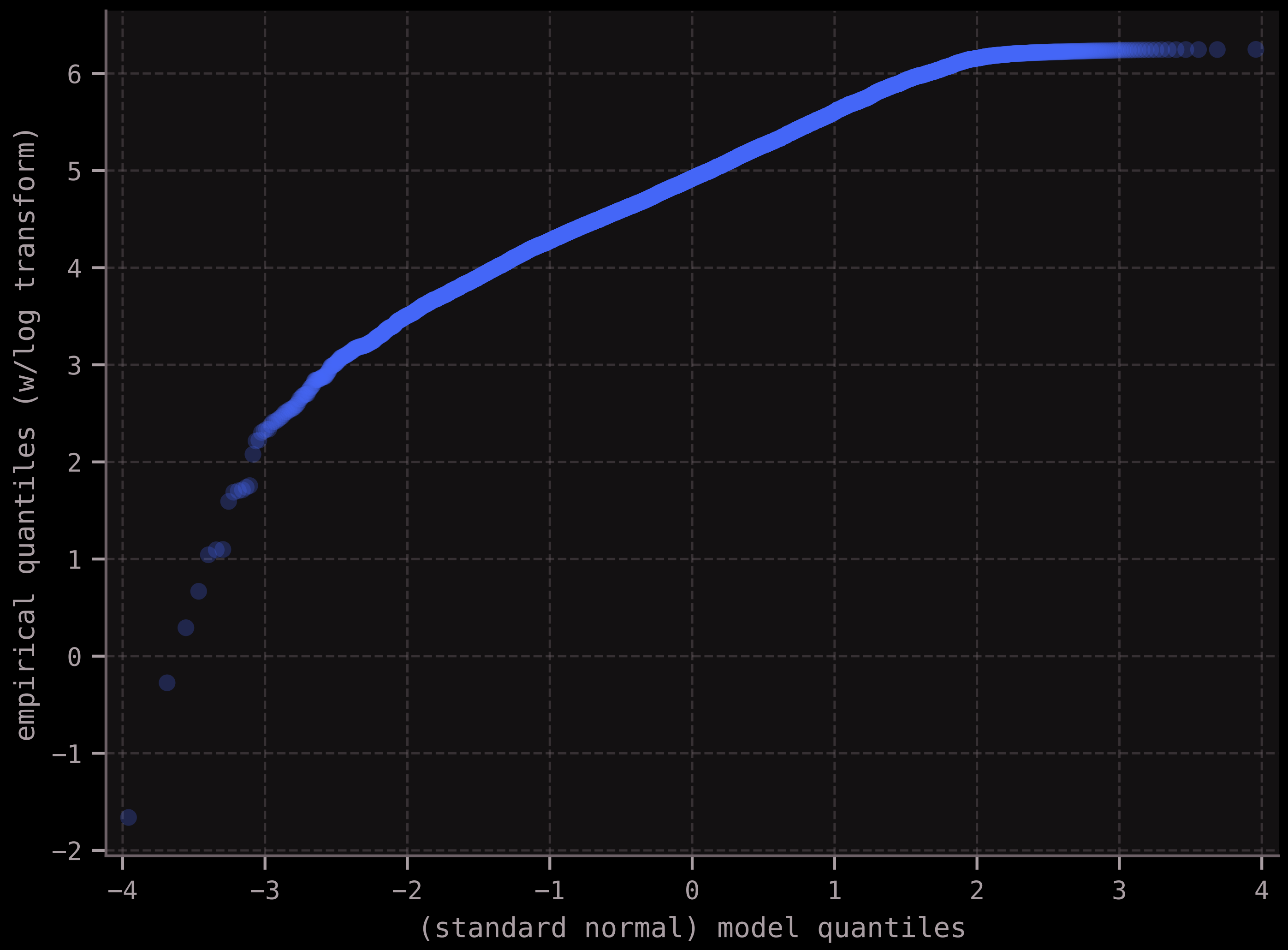


600
500
400
300
200
100
0
−100
−200
empirical quantiles
−200
−100
0
100
200
300
400
500
600
(normal) model quantiles

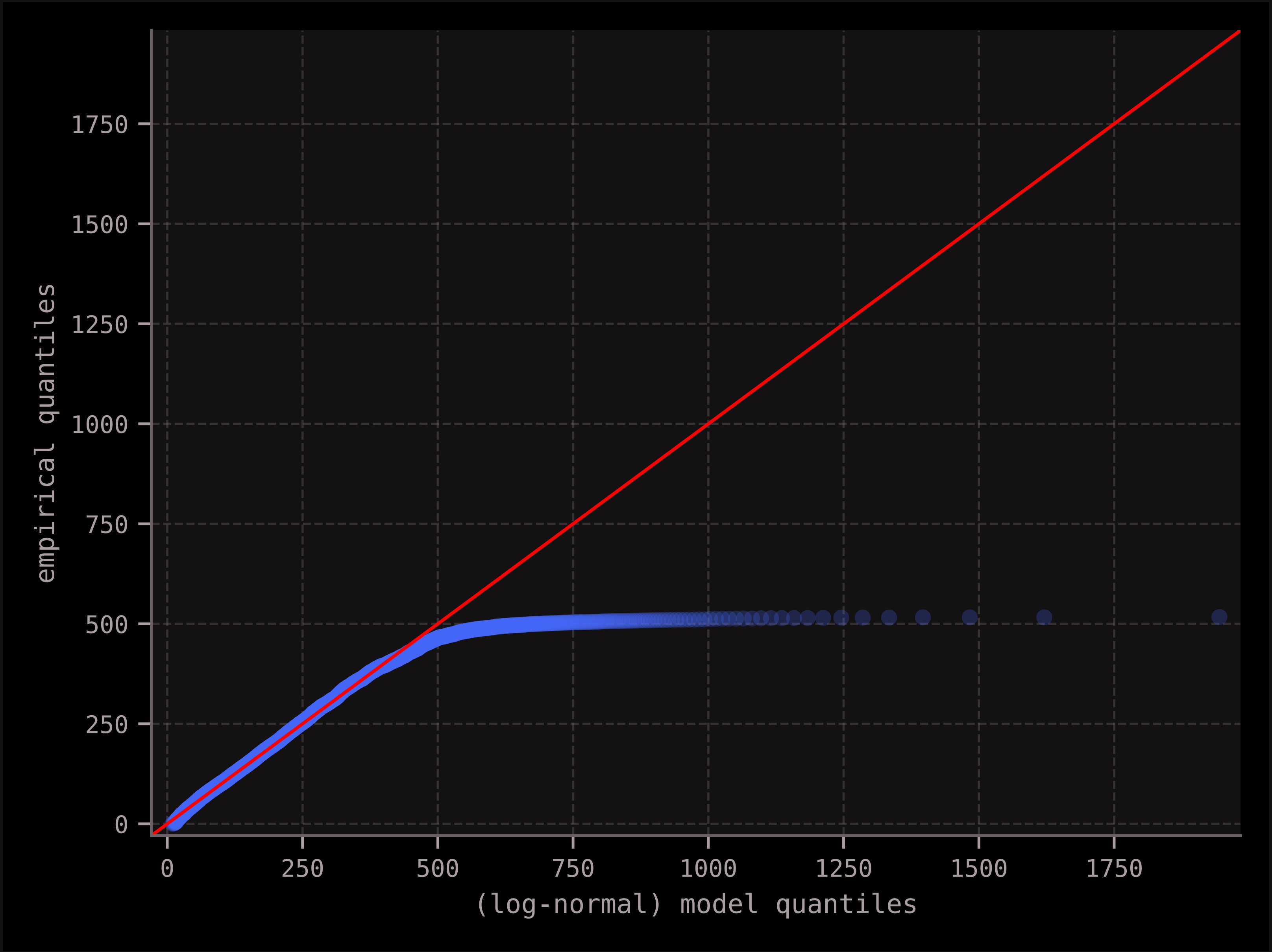


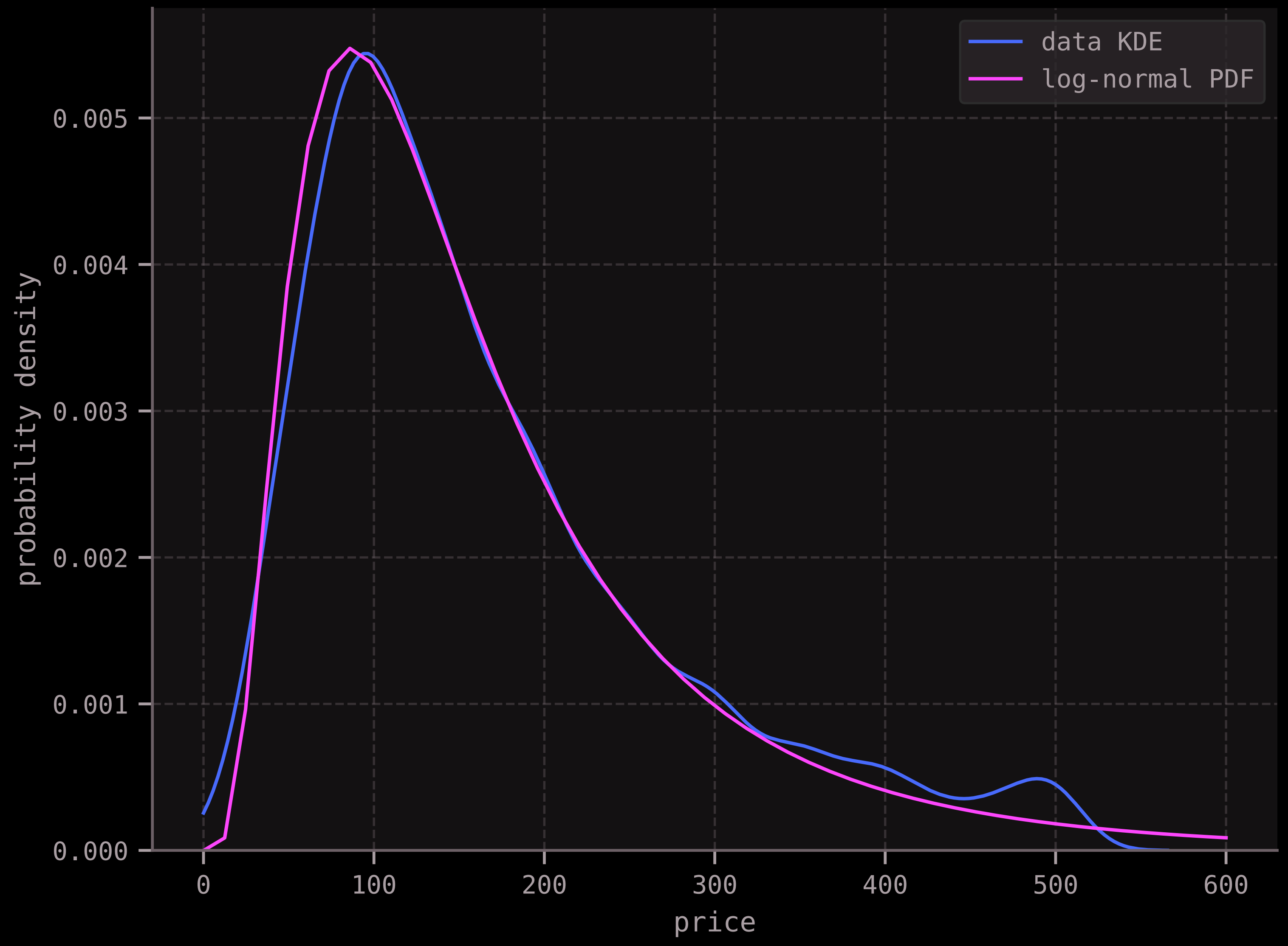
with diagonal line
empirical quantiles
0
100
200
300
400
500
0
100
200
300
400
500
(standard normal) model quantiles
without diagonal line
empirical quantiles
−4
−3
−2
−1
0
1
2
3
4
(standard normal) model quantiles

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 $\mathcal{N}(b, a^2)$.

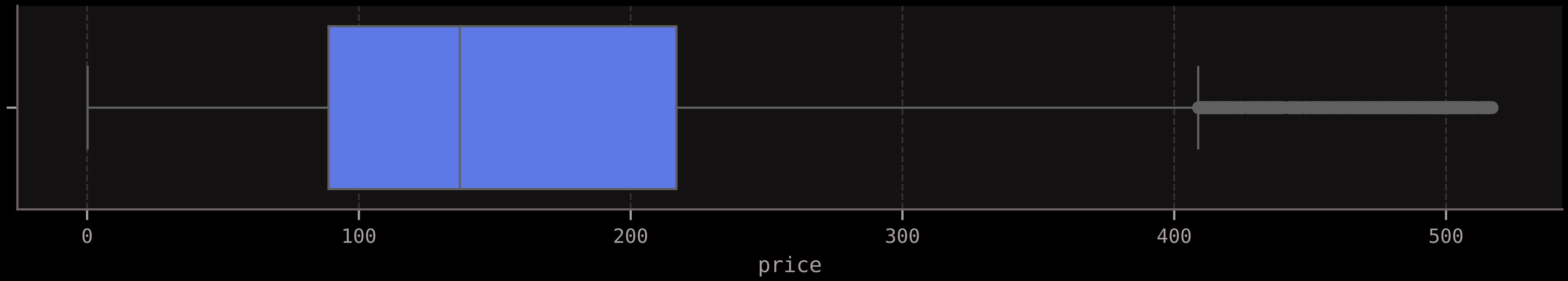


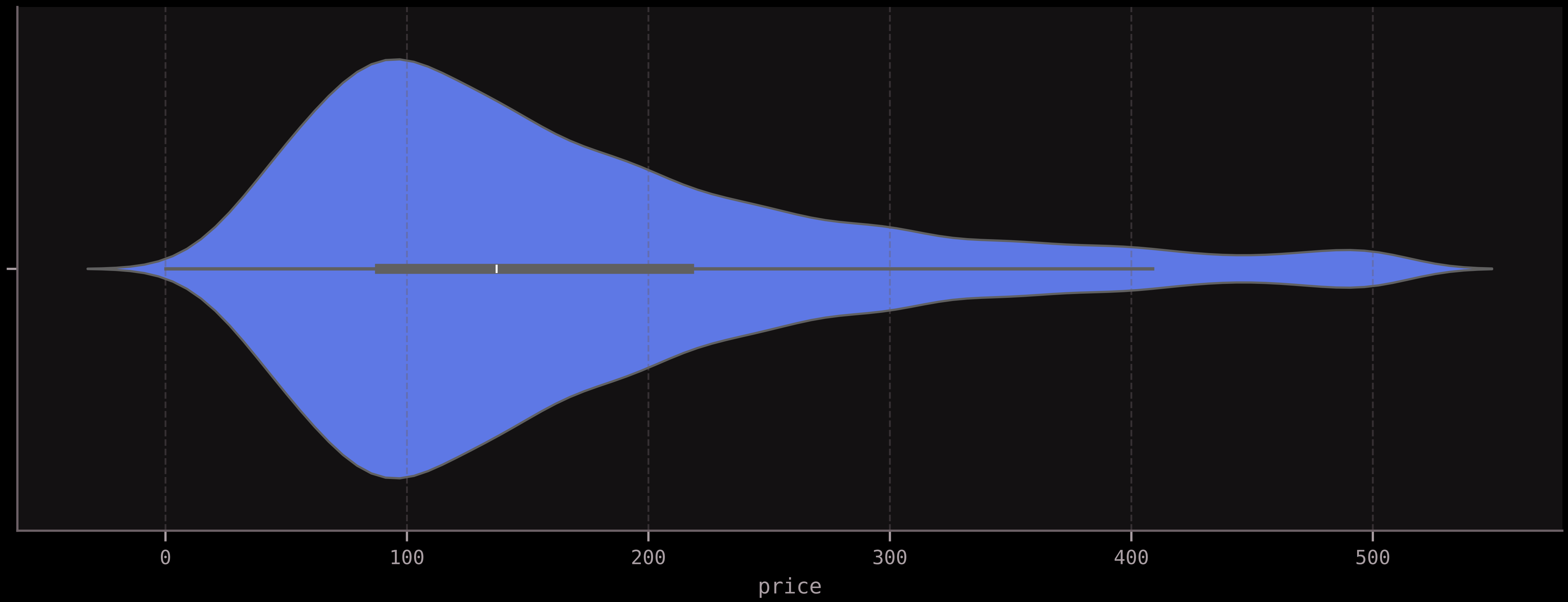
empirical quantiles (w/log transform)
6
5
4
3
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4
(standard normal) model quantiles


empirical quantiles
(log-normal) model quantiles
0
250
500
750
1000
1250
1500
1750

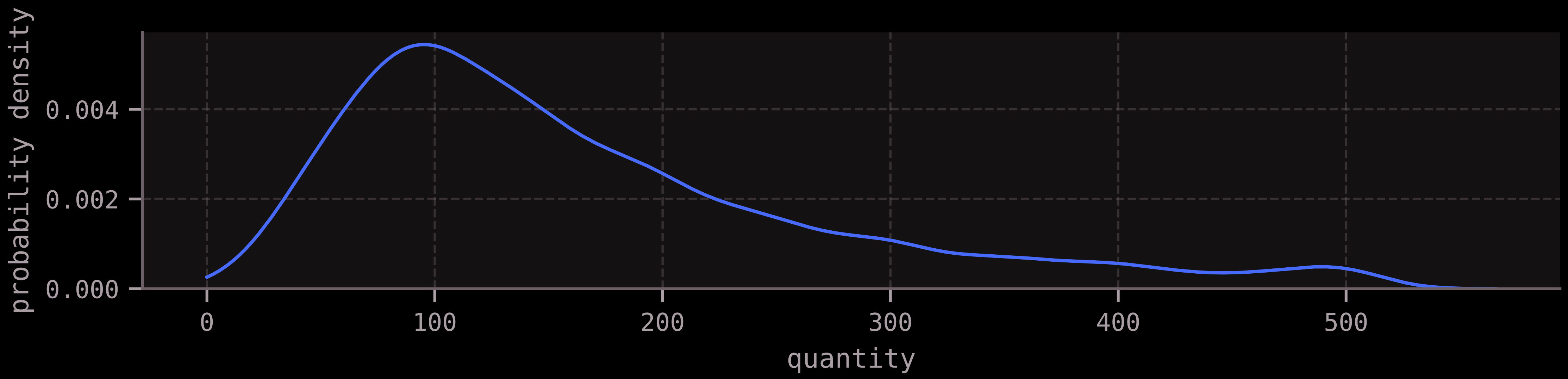

data KDE
log-normal PDF
probability density
0.000
0.001
0.002
0.003
0.004
0.005
0
100
200
300
400
500
600
price

6.7. Box plots and violin plots


0
100
200
300
400
500
price



0
100
200
300
400
500
price


probability density
0.004
0.002
0.000
0
100
200
300
400
500
quantity