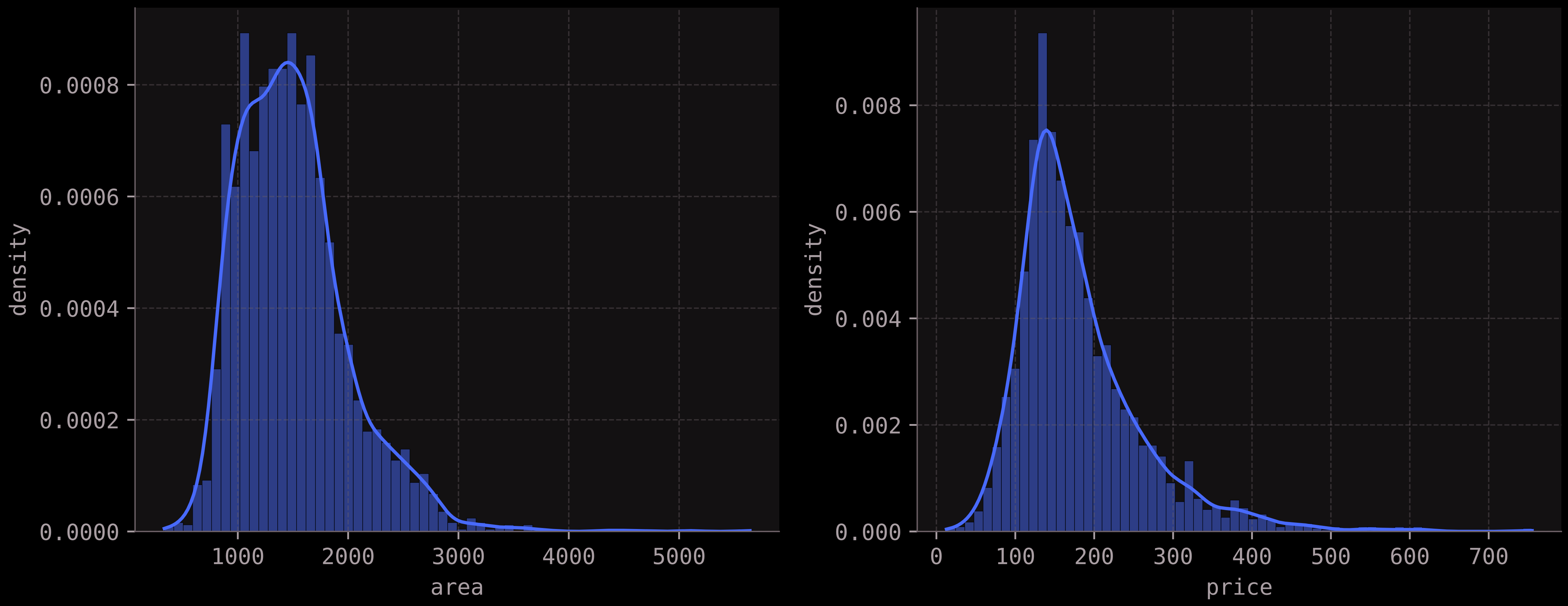
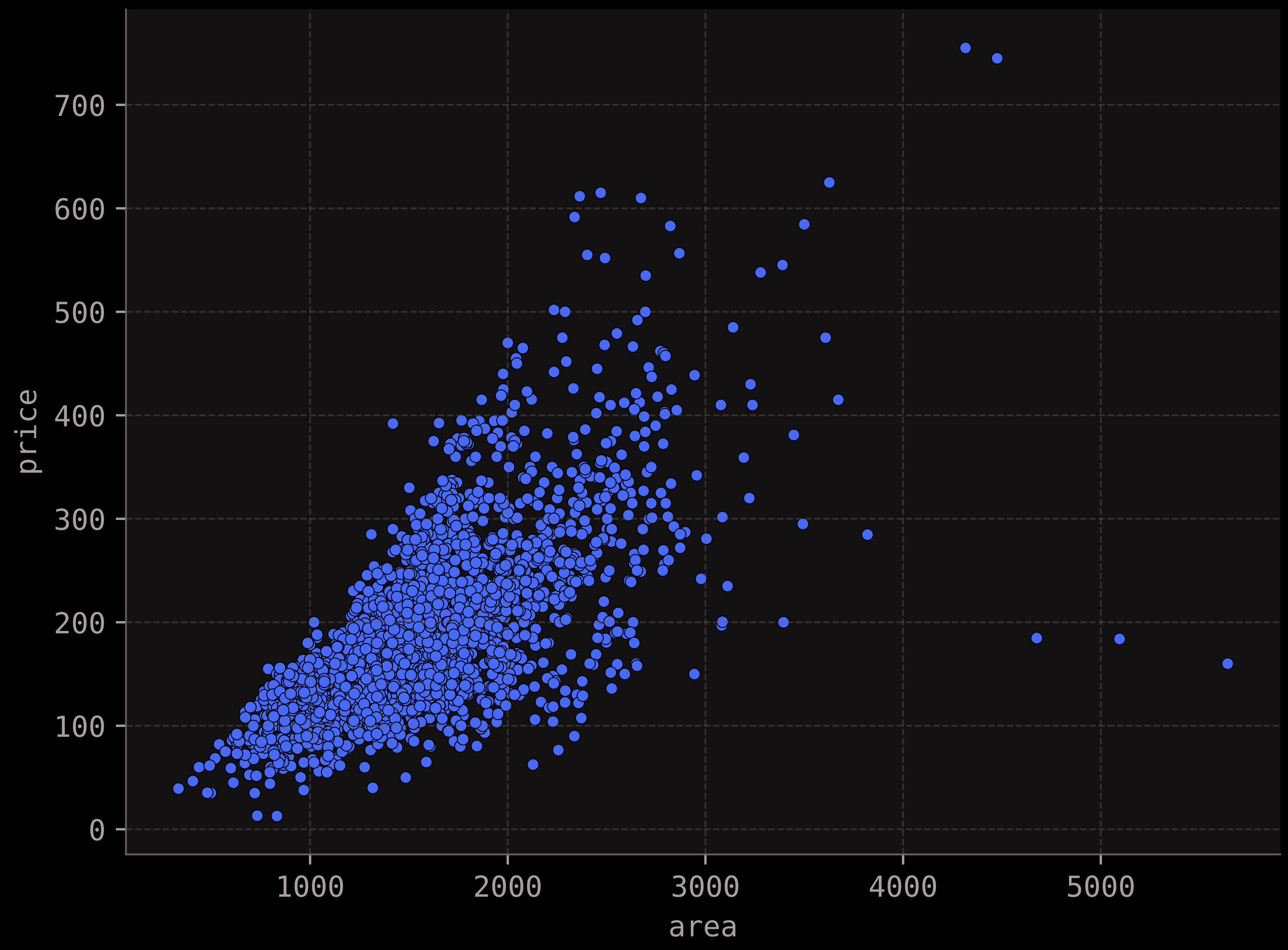
7. Random vectors

7.1. Motivation



0.0008
0.0006
0.0004
0.0002
0.0000
density
1000
2000
3000
4000
5000
area
0.008
0.006
0.004
0.002
0.000
density
0
100
200
300
400
500
600
700
price



price
700
600
500
400
300
200
100
0
1000
2000
3000
4000
5000
area

7.2. 2-dimensional random vectors

Definition 7.1

Let S be a probability space. A 2-*dimensional random vector* is a function

$$\mathbf{X} : S \to \mathbb{R}^2.$$

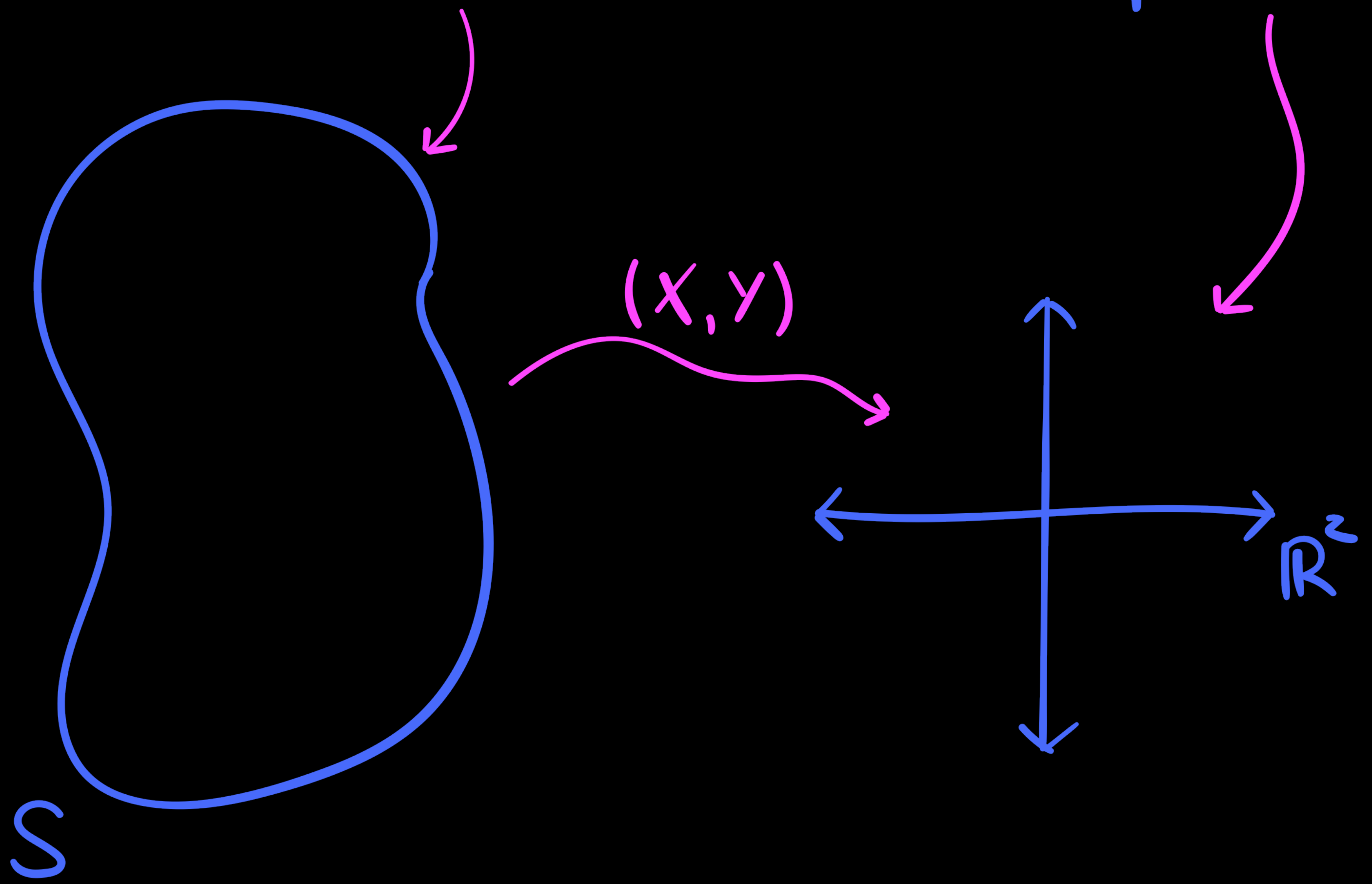
Thus, we may write $\mathbf{X}(s) = (X_1(s), X_2(s))$ for each sample point $s \in S$, where

$$X_1 : S \to \mathbb{R} \quad \text{and} \quad X_2 : S \to \mathbb{R}$$

are random variables. When we do so, the random variables X_1 and X_2 are called the *components* of the random vector $\mathbf{X}$.

P measures probability here

P_{XY} measures probability here

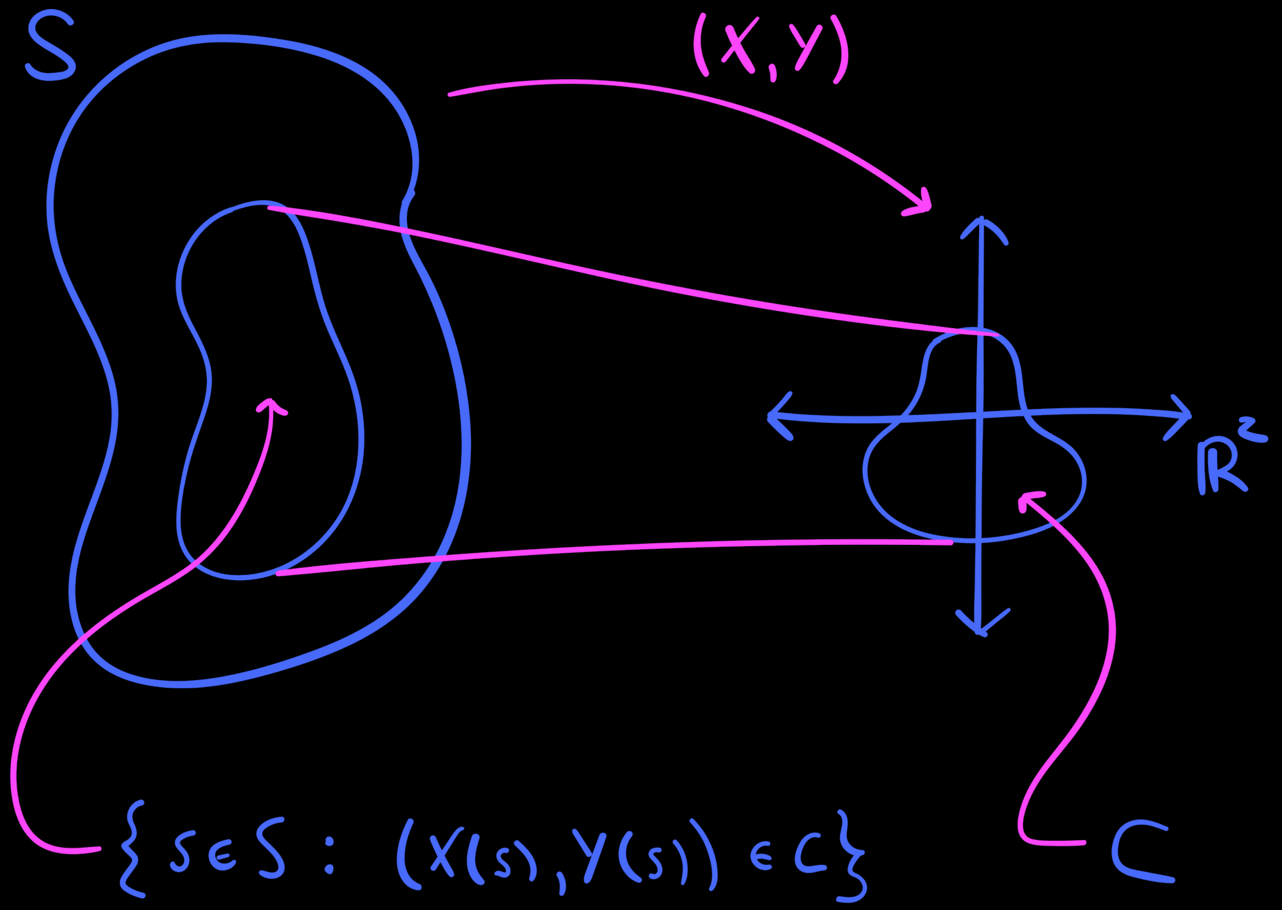


Definition 7.2

Let $(X, Y) : S \to \mathbb{R}^2$ be a 2-dimensional random vector on a probability space S with probability measure P. We define the *probability measure* of (X, Y), denoted P_{XY}, via the formula

$$P_{XY}(C) = P(\{s \in S : (X(s), Y(s)) \in C\}), \tag{7.1}$$

for all events $C \subset \mathbb{R}^2$. The probability measure P_{XY} is also called the *joint distribution* or the *bivariate distribution* of X and Y.



these have equal probabilities by definition!

Problem Prompt

Do problem 1 on the worksheet.

Definition 7.3

Let (X, Y) be a 2-dimensional random vector.

- We shall say (X, Y) is *discrete*, or that X and Y are *jointly discrete*, if the joint probability distribution P_{XY} is discrete. In other words, we require that there exists a *joint probability mass function* $p(x, y)$ such that

  $$P((X, Y) \in C) = \sum_{(x,y)\in C} p(x, y)$$

  for all events $C \subset \mathbb{R}^2$.
- We shall say (X, Y) is *continuous*, or that X and Y are *jointly continuous*, if the joint probability distribution P_{XY} is continuous. In other words, we require that there exists a *joint probability density function* $f(x, y)$ such that

  $$P\left((X, Y) \in C\right) = \iint_C f(x, y) \, \mathrm{d}x\mathrm{d}y$$

  for all events $C \subset \mathbb{R}^2$.

Theorem 7.1

Let (X, Y) be a 2-dimensional random vector.

1. The random vector (X, Y) is discrete if and only if both X and Y are discrete.
2. If (X, Y) is continuous, then X and Y are both continuous. However, it does *not* nececessarily follow that if both X and Y are continuous, then so too is (X, Y).

Problem Prompt

Do problems 2-4 on the worksheet.

7.3. Bivariate distribution functions

Definition 7.4

Let (X, Y) be a 2-dimensional random vector. The *distribution function* of (X, Y) is the function $F : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F(x, y) = P(X \leq x, Y \leq y).$$

In particular:

1. If (X, Y) is discrete with probability mass function $p(x, y)$, then

$$F(x, y) = \sum_{x^\star \leq x,\, y^\star \leq y} p(x^\star, y^\star).$$

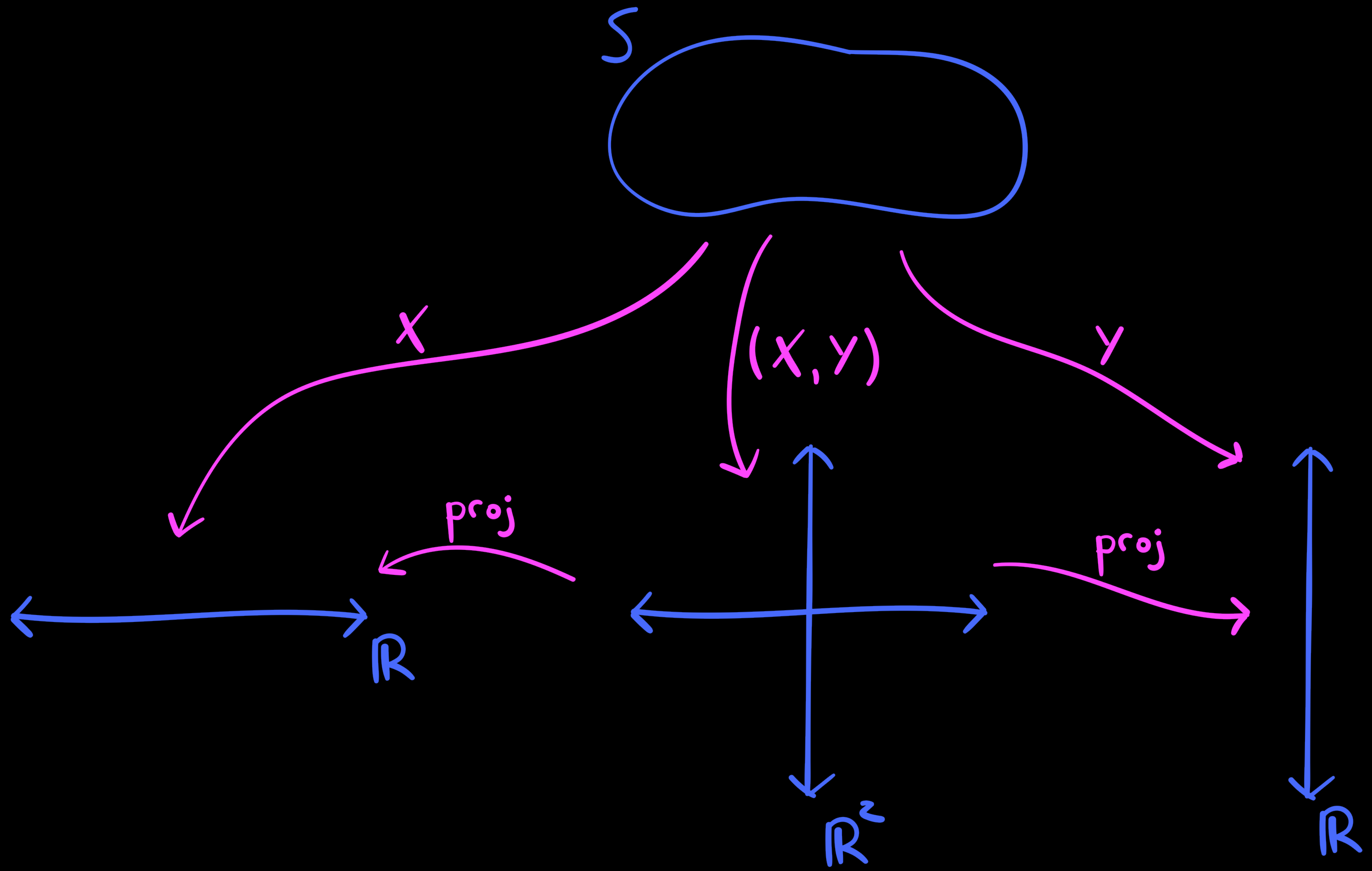
2. If (X, Y) is continuous with probability density function $f(x, y)$, then

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x^\star, y^\star)\, \mathrm{d}x^\star \mathrm{d}y^\star.$$

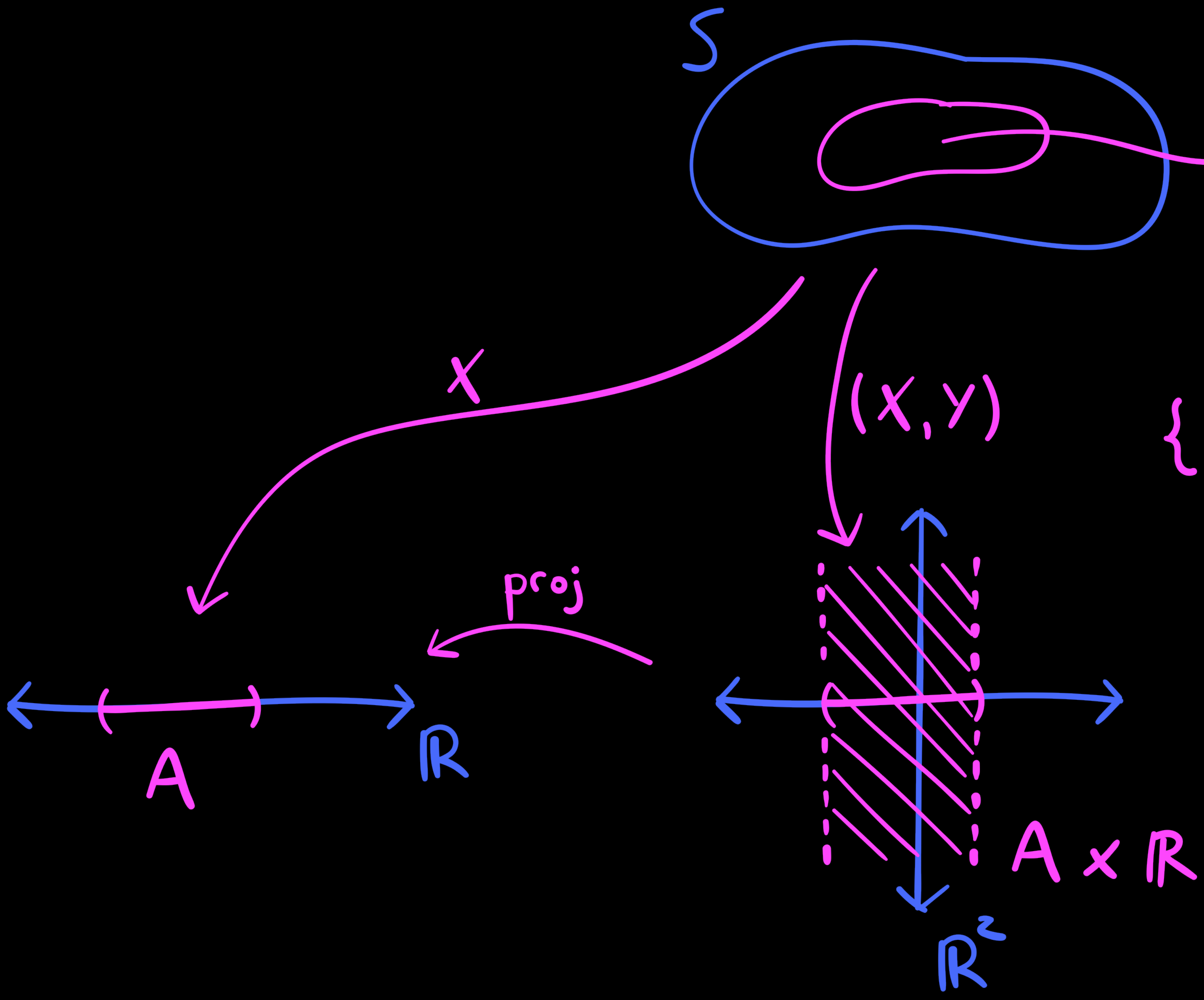
Problem Prompt

Do problem 5 on the worksheet.

7.4. Marginal distributions



S
X
(X, Y)
Y
proj
proj
ℝ
ℝ²
ℝ



this is equal to both

$$\{ s \in S : X(s) \in A \}$$

and

$$\{ s \in S : (X(s), Y(s)) \in A \times \mathbb{R} \}$$

Theorem 7.2

Let (X, Y) be a 2-dimensional random vector with induced probability measure P_{XY}. Then the measures P_X and P_Y may be obtained via the formulas

$$P_X(A) = P_{XY}(A \times \mathbb{R}) \quad \text{and} \quad P_Y(B) = P_{XY}(\mathbb{R} \times B)$$

for all events $A, B \subset \mathbb{R}$. In particular:

1. If (X, Y) is discrete with probability mass function $p(x, y)$, then

$$P(X \in A) = \sum_{x \in A,\, y \in \mathbb{R}} p(x, y) \quad \text{and} \quad P(Y \in B) = \sum_{y \in B,\, x \in \mathbb{R}} p(x, y).$$

2. If (X, Y) is continuous with probability density function $f(x, y)$, then

$$P(X \in A) = \int_A \int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}x\mathrm{d}y$$

and

$$P(Y \in B) = \int_B \int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}y\mathrm{d}x.$$

Definition 7.5

Let (X, Y) be a 2-dimensional random vector. Then the distributions P_X and P_Y are called the *marginal distributions* of (X, Y).

Theorem 7.3

Let (X, Y) be a 2-dimensional random vector.

1. If (X, Y) is discrete with probability mass function $p(x, y)$, then both X and Y are discrete with probability mass functions given by

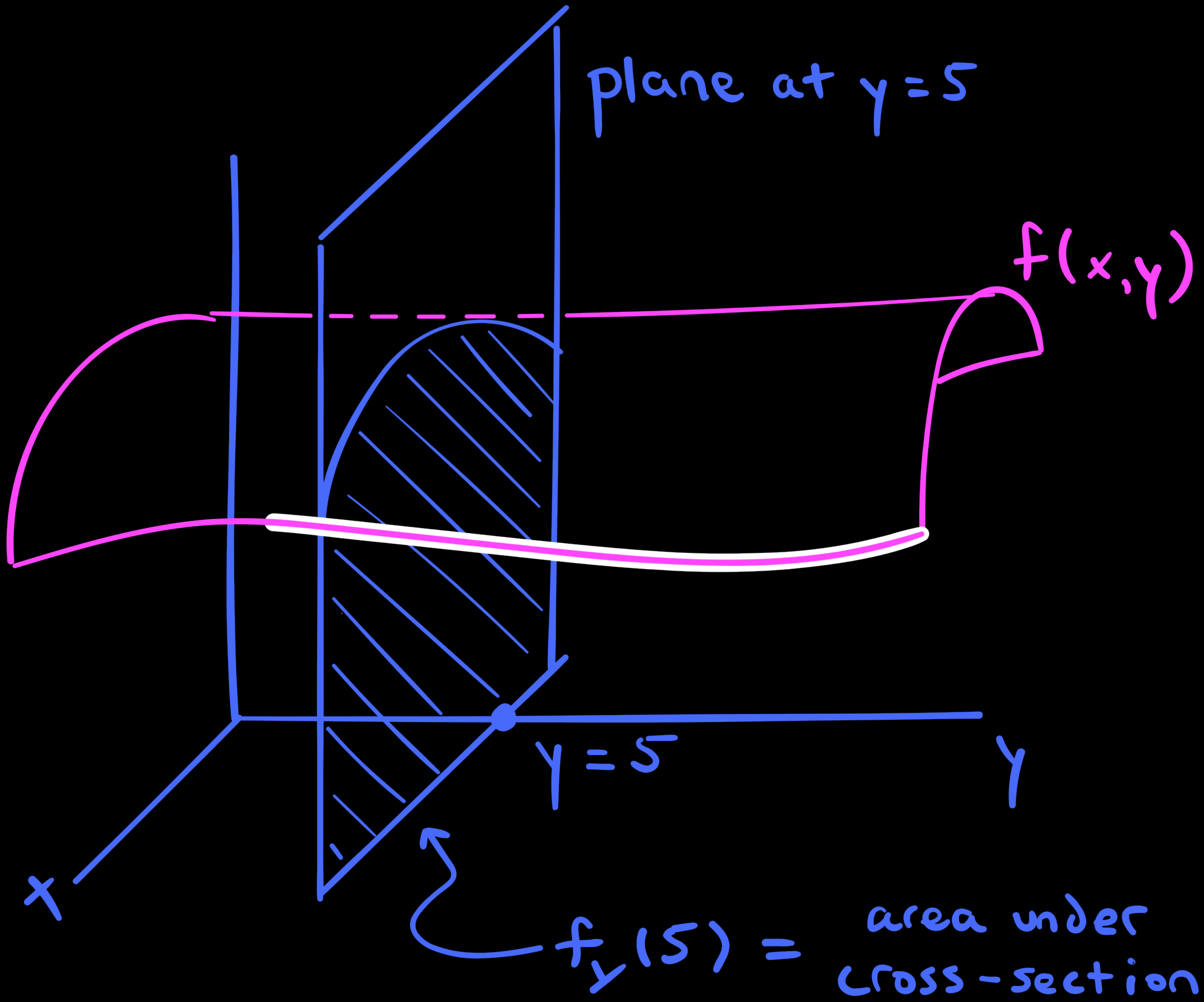
$$p_X(x) = \sum_{y \in \mathbb{R}} p(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x \in \mathbb{R}} p(x, y).$$

2. If (X, Y) is continuous with probability density function $f(x, y)$, then both X and Y are continuous with probability density functions given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x. \tag{7.5}$$

Tip

1. To obtain the marginal density $f_X(x)$ from the joint density $f(x, y)$, we "integrate out" the dependence of $f(x, y)$ on y. Likewise for obtaining $f_Y(y)$ from $f(x, y)$.
2. To obtain the marginal mass $p_X(x)$ from the joint mass $p(x, y)$, we "sum out" the dependence of $p(x, y)$ on y. Likewise for obtaining $p_Y(y)$ from $p(x, y)$.



plane at y=5
f(x,y)
x
y=5
y
fy(5) = area under cross-section

Problem Prompt

Do problems 6 and 7 on the worksheet.

7.5. Bivariate empirical distributions

Definition 7.6

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)$ be a sequence of 2-dimensional random vectors, all defined on the same probability space.

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

Provided that the sequence is a bivariate random sample, an *observed bivariate random sample*, or a *bivariate dataset*, is a sequence of pairs of real numbers

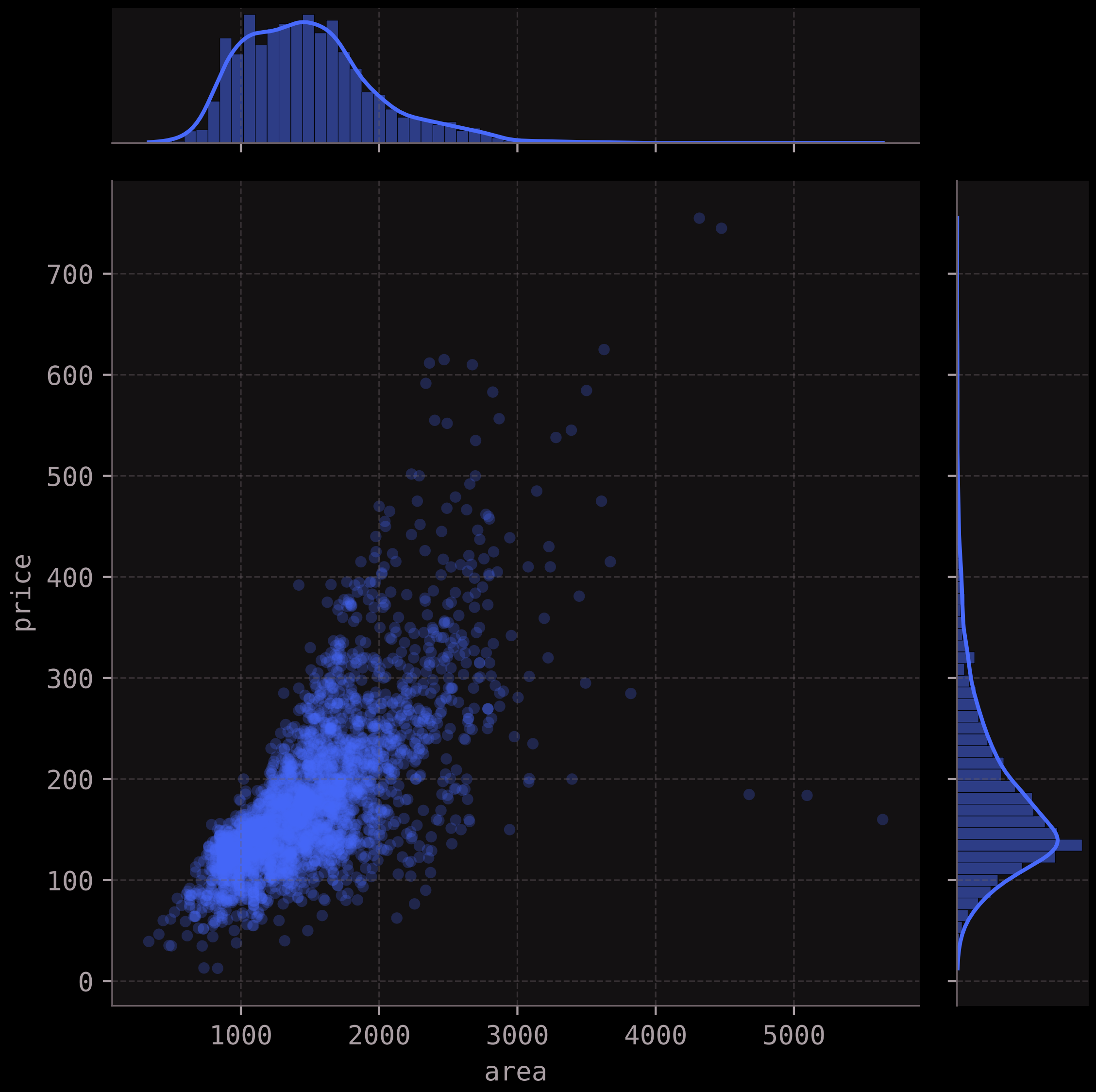
$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$$

where (x_i, y_i) is an observation of (X_i, Y_i).

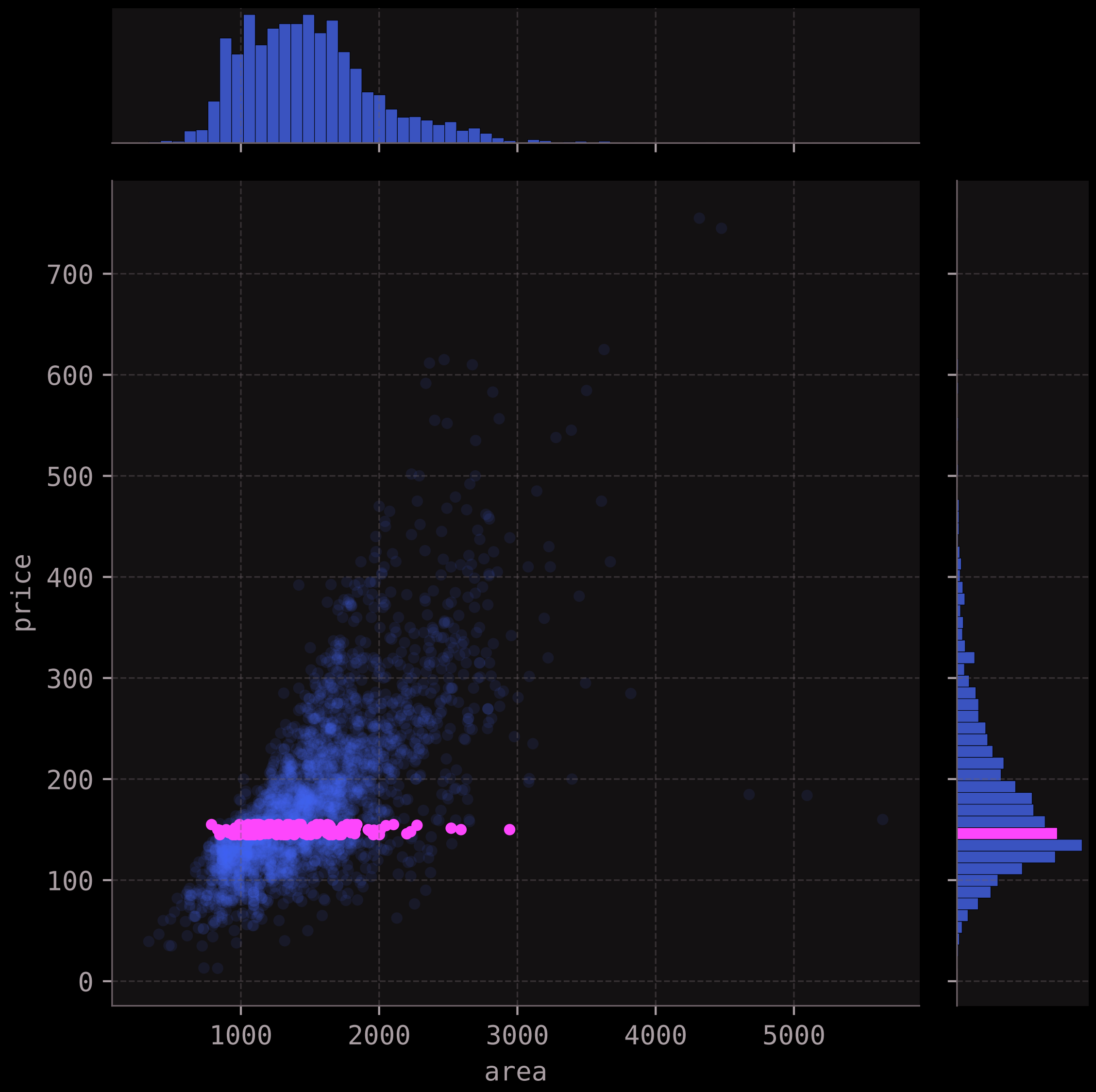
Definition 7.7

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ be an observed bivariate random sample, i.e., a bivariate dataset. The *empirical distribution* of the dataset is the discrete probability measure on $\mathbb{R}^2$ with joint probability mass function

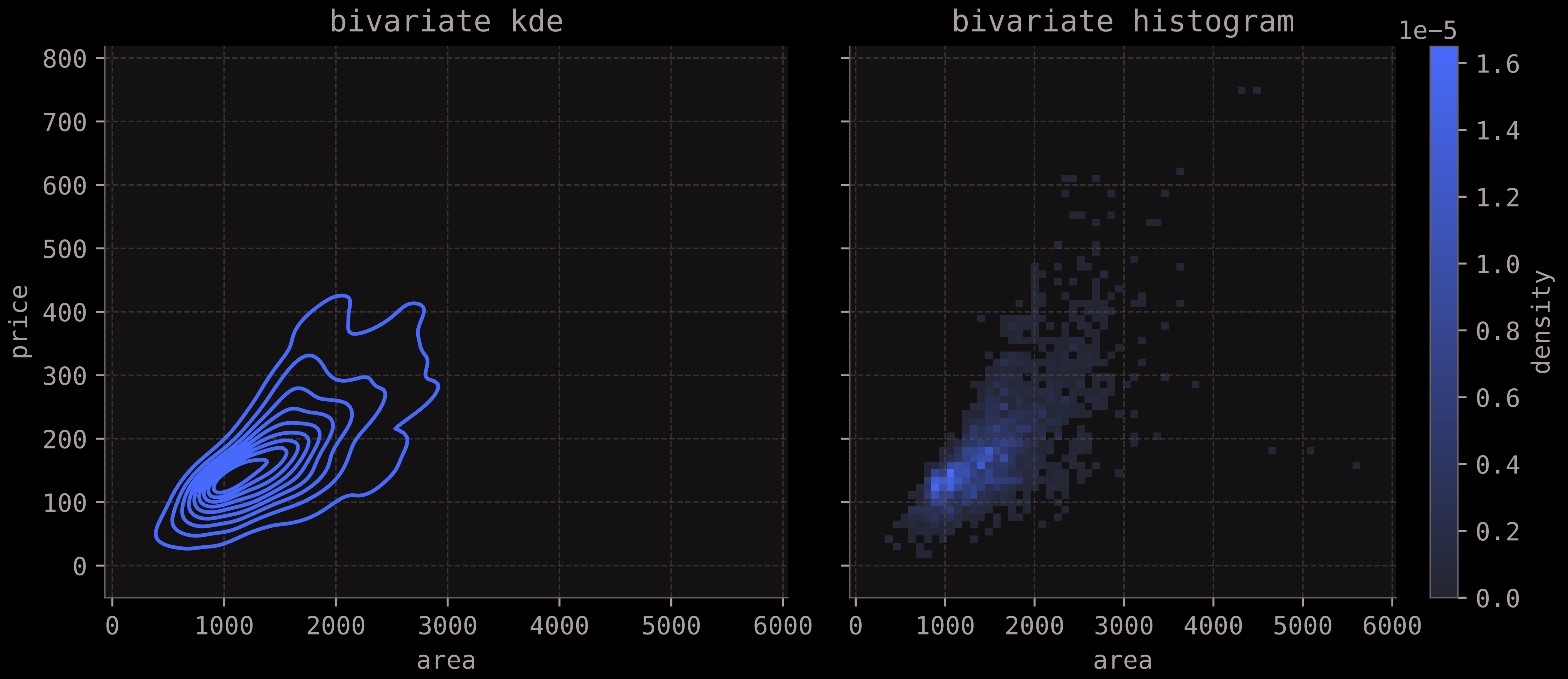
$$p(x,y) = \frac{\text{number of data points } (x_i, y_i) \text{ that match } (x,y)}{m}.$$



price
700
600
500
400
300
200
100
0
1000
2000
3000
4000
5000
area



price
0
100
200
300
400
500
600
700
area
1000
2000
3000
4000
5000


bivariate kde
bivariate histogram
price
area
area
density
1e−5
0
100
200
300
400
500
600
700
800
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6