Statistical Inference and Machine Learning in Earth Science
SIMLES

Module 1
Basics

Lecture2
Random Variables & Distributions

Cristi Proistosescu

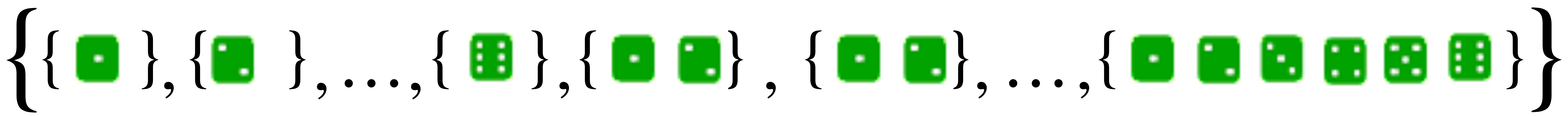
Probability space $(\mathcal{S}, \mathcal{E}, P)$

In order to define a random process and work with probabilities, we need to construct a probability space.

Sample space: (Set of all possible outcomes/rolls)

$\mathcal{S} = \{$  $\}$

Event space: (Set of all subsets of the sample space)

$\mathcal{E} =$ 

Valid Probability: P

- $P > 0$, $P(\phi) = 0$, $P(\mathcal{S}) = 1$
- If E_1, E_2 are mutual exclusive $(E_1 \cap E_2 = \phi)$, the probability of E_1 OR E_2 is:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Random variable

A random variable X is a function defined on the sample space,
That associates a number for outcome and event:

$$X : \mathcal{S} \rightarrow \mathbb{R}$$

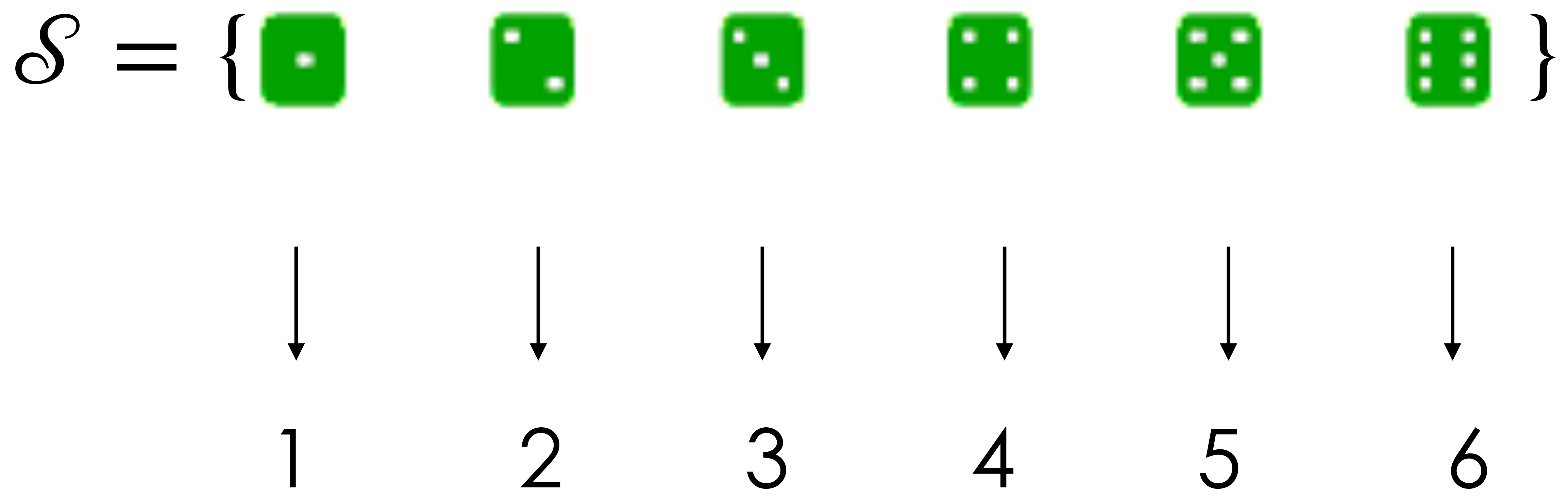
X: random variable

x : a specific value the random variable takes in the real numbers

Uniform discrete random variable

$$X : \mathcal{S} \rightarrow \{1,2,3,4,5,6\}$$

X: value of each die face



Bernoulli Random variable

$$X : \mathcal{S} \to \{0,1\}$$

$$X = 0 \text{ if roll } \leq 3$$
$$= 1 \text{ if roll } > 3$$

$$p = P(X = 1)$$

$$p = 0.5$$

$\mathcal{S} = \{$  $\}$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

0 0 0 1 1 1

Same sample space, different RV

Uniform discrete

$$X : \{1,2,3,4,5,6\} \to \{1,2,3,4,5,6\}$$

Bernoulli

$$X : \{1,2,3,4,5,6\} \to \{0,1\}$$

Bernoulli Random variable

$$X : \mathcal{S} \to \{0,1\}$$

$X = 0$ if heads

$= 1$ if tails

$$p = P(X = 1)$$

$$p = 0.5$$

0 1

Bernoulli Random variable

$$X : \mathcal{S} \to \{0,1\}$$

$X = 0$ if roll ≤ 2

$\;\;\;= 1$ if roll > 2

$$p = P(X = 1)$$

$$p = 2/3$$

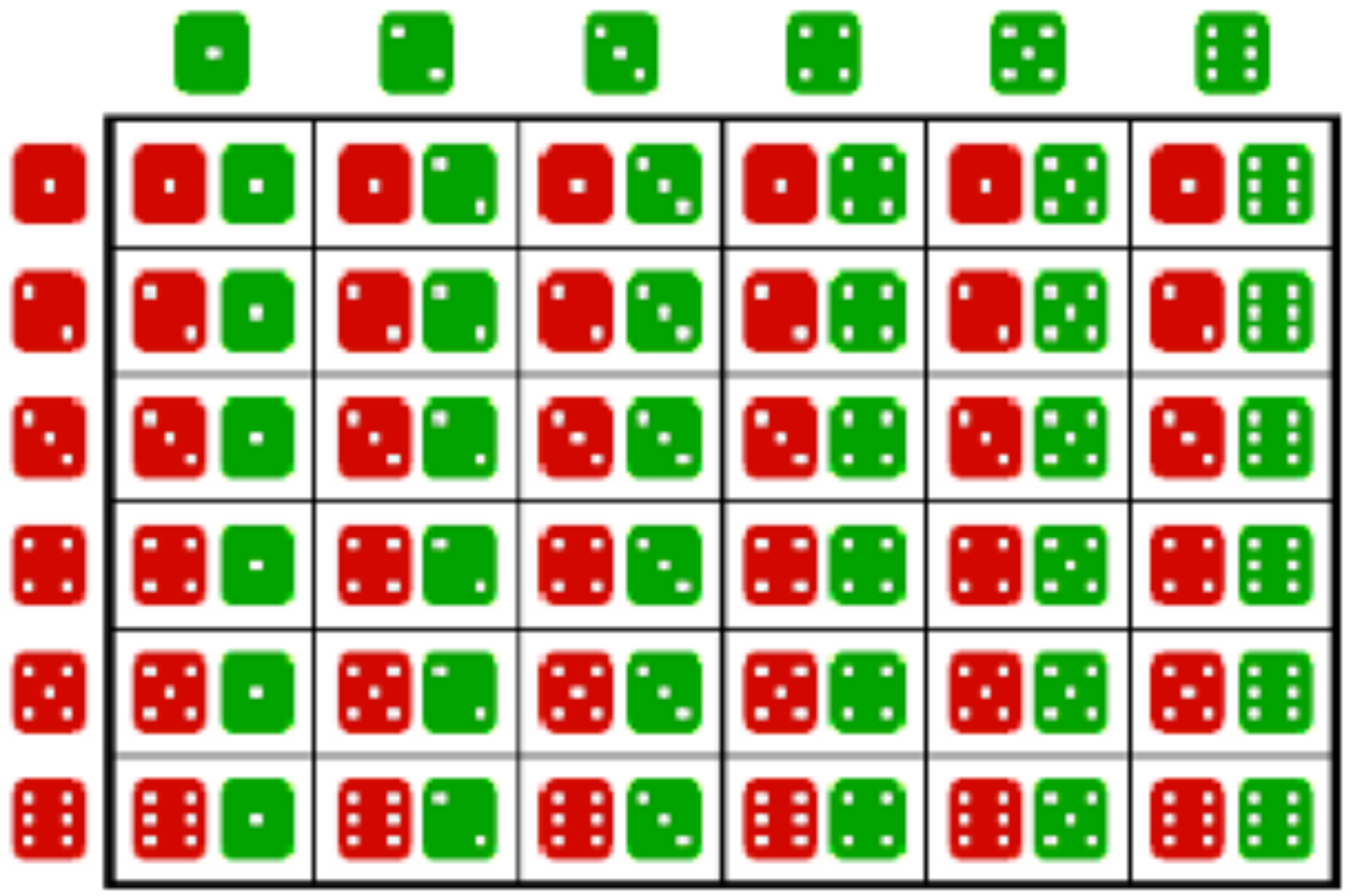
$\mathcal{S} = \{$ ⚀ ⚁ ⚂ ⚃ ⚄ ⚅ $\}$

$$\begin{array}{cccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 0 & 0 & 1 & 1 & 1 & 1 \end{array}$$

Sum of two dice

$$X : \mathcal{S} \to \{2,\ldots,12\}$$

$\mathcal{S}$:



$\to$

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

$P = 1/36$ for each event in sample space

Distributions: probability mass function

Random variables are characterized (and defined) by their distribution

$$X : \mathcal{S} \rightarrow \{x_i\}$$

Probability mass function (pmf) of X:

$$p(x_i) = P(X = x_i)$$

Distributions: probability mass function

Uniform discrete: $p(x_i) = 1/6$

Bernoulli: $p(0) = 1 - p;\ p(1) = p$

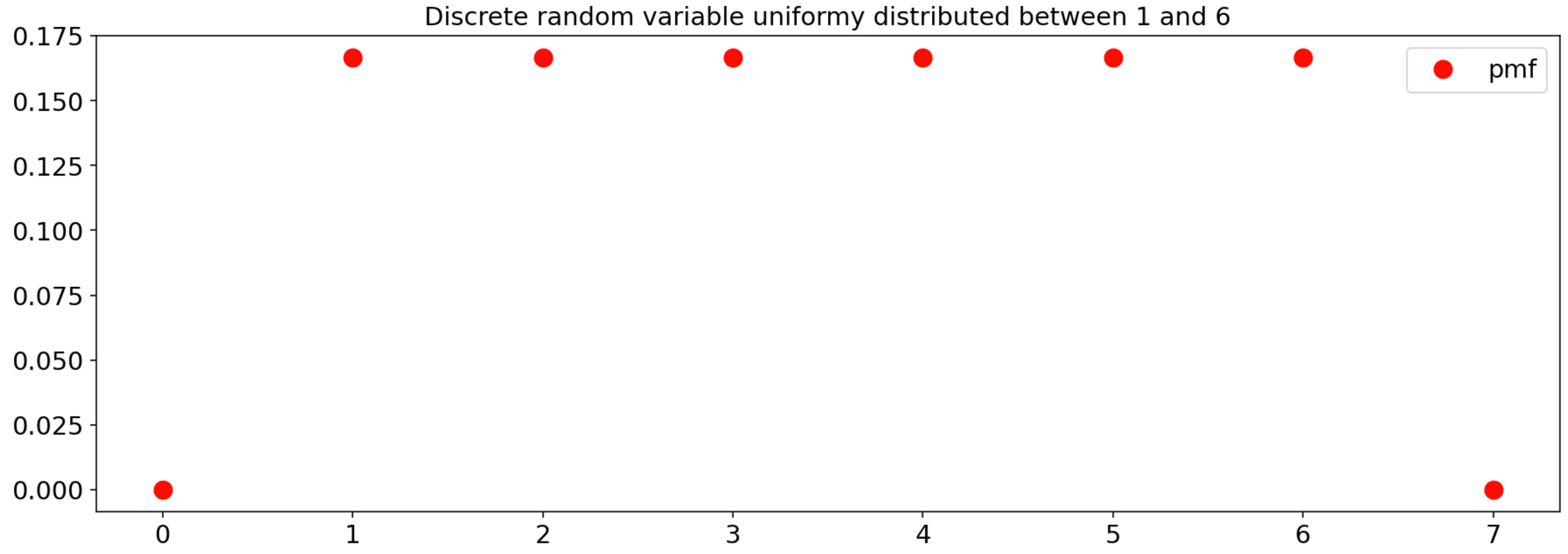
Sum of two dice:

- $p(2) = 1/36$
- $p(3) = 2/36$
- …
- $p(12) = 1/36$

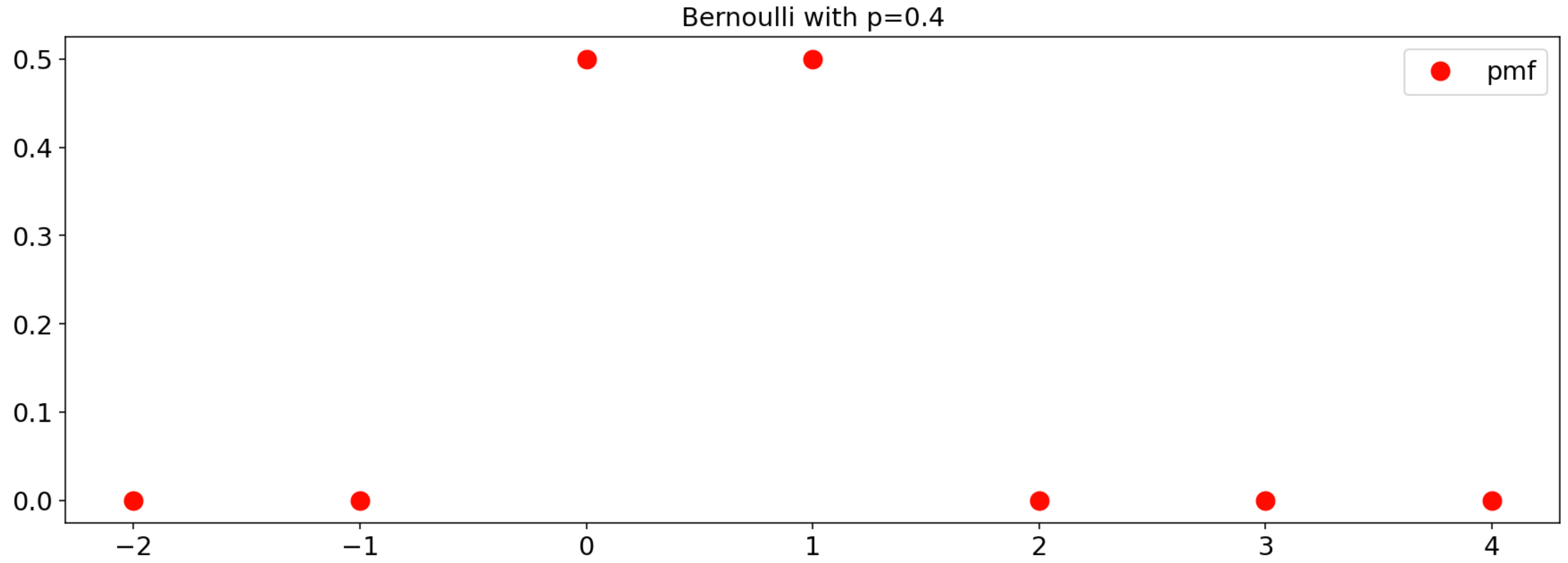
	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

Distributions: probability mass function

Uniform discrete: $p(x_i) = 1/6$

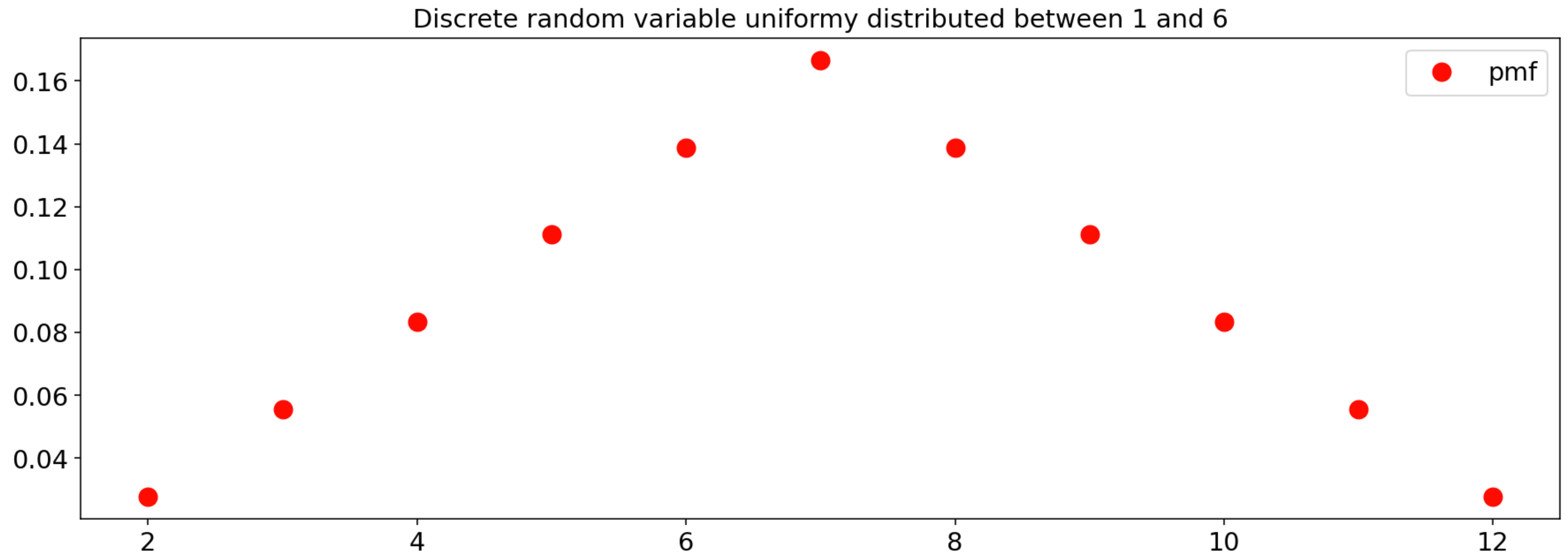


Bernoulli: $p(0) = 1 - p;\ p(1) = p$



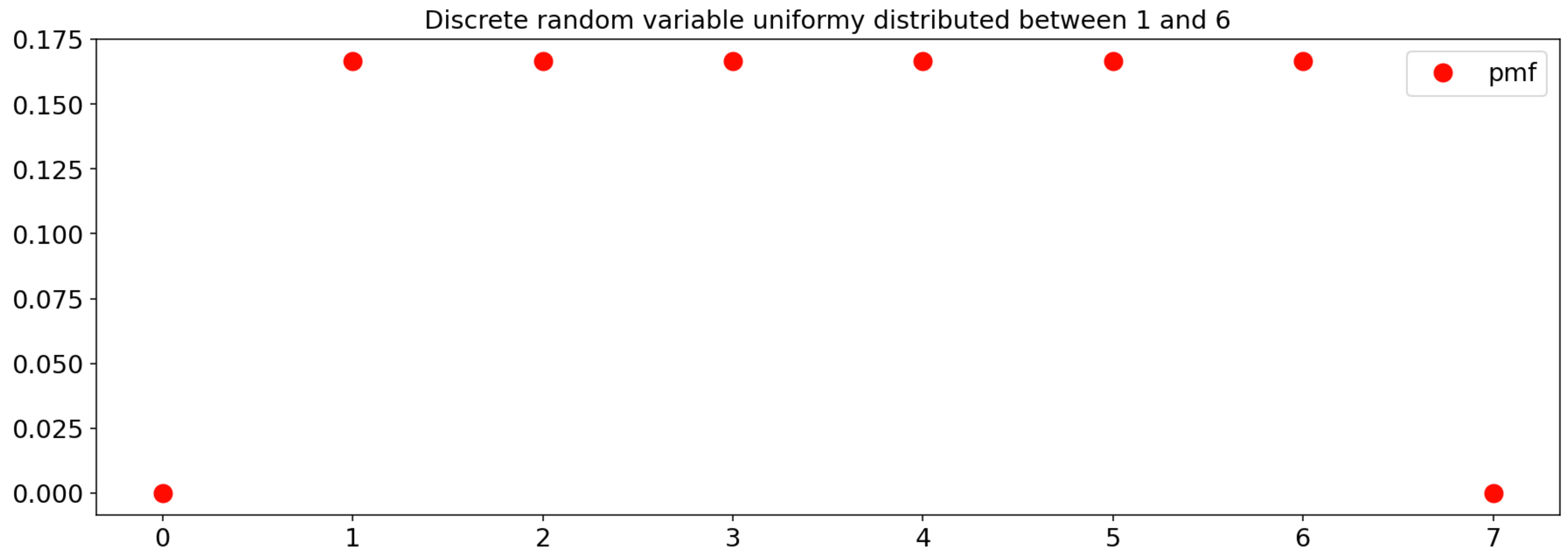
Sum of two dice:

- $p(2) = 1/36$
- $p(3) = 2/36$
- $\ldots$
- $p(12) = 1/36$

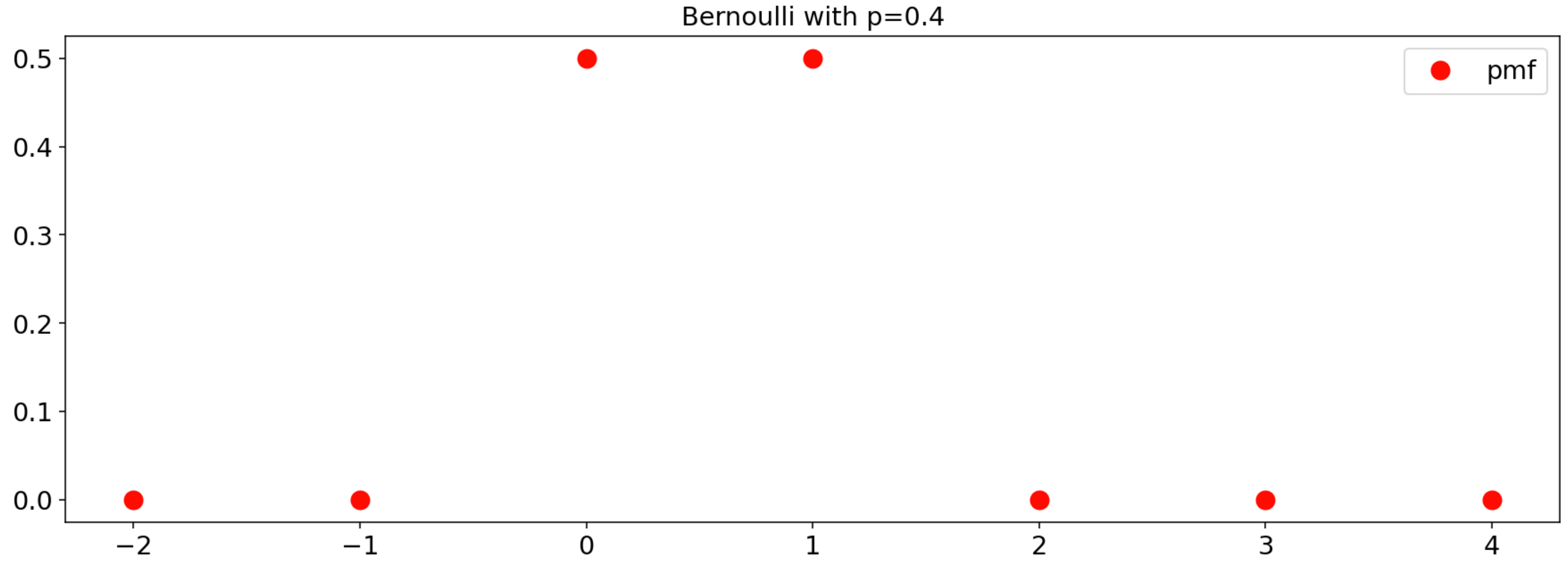


Distributions: probability mass function

Uniform discrete: $p(x_i) = 1/6$

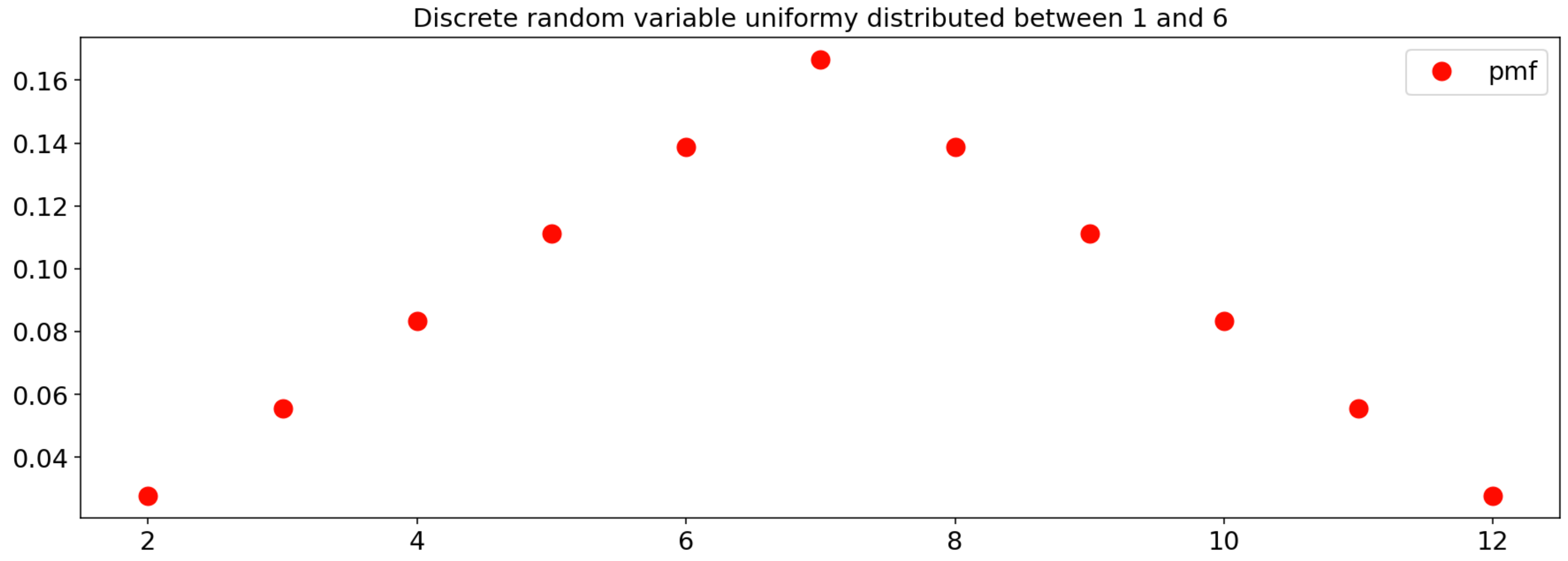


Bernoulli: $p(0) = 1 - p;\ p(1) = p$



Sum of two dice:

- $p(2) = 1/36$
- $p(3) = 2/36$
- $\ldots$
- $p(12) = 1/36$



Probability ↔ Statistics

Probability

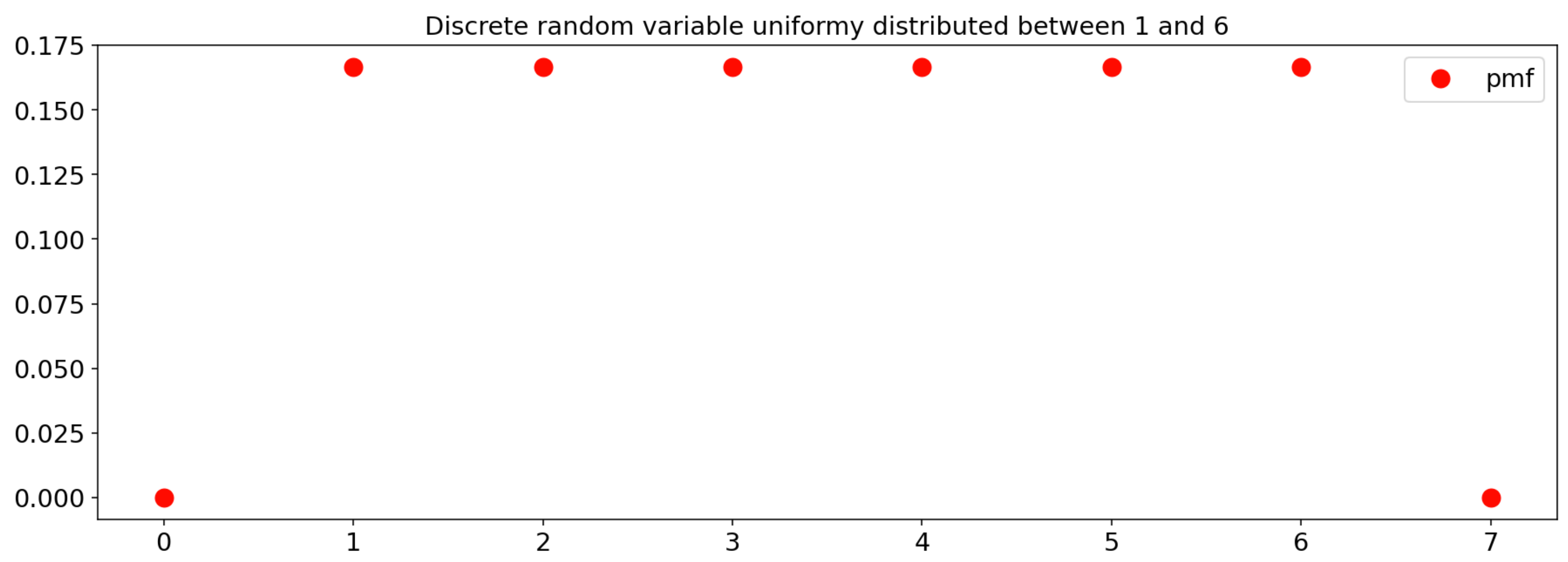
Process

↓

Distribution $p(x_i) = 1/6$

↓

Data



Probability ↔ Statistics

Probability

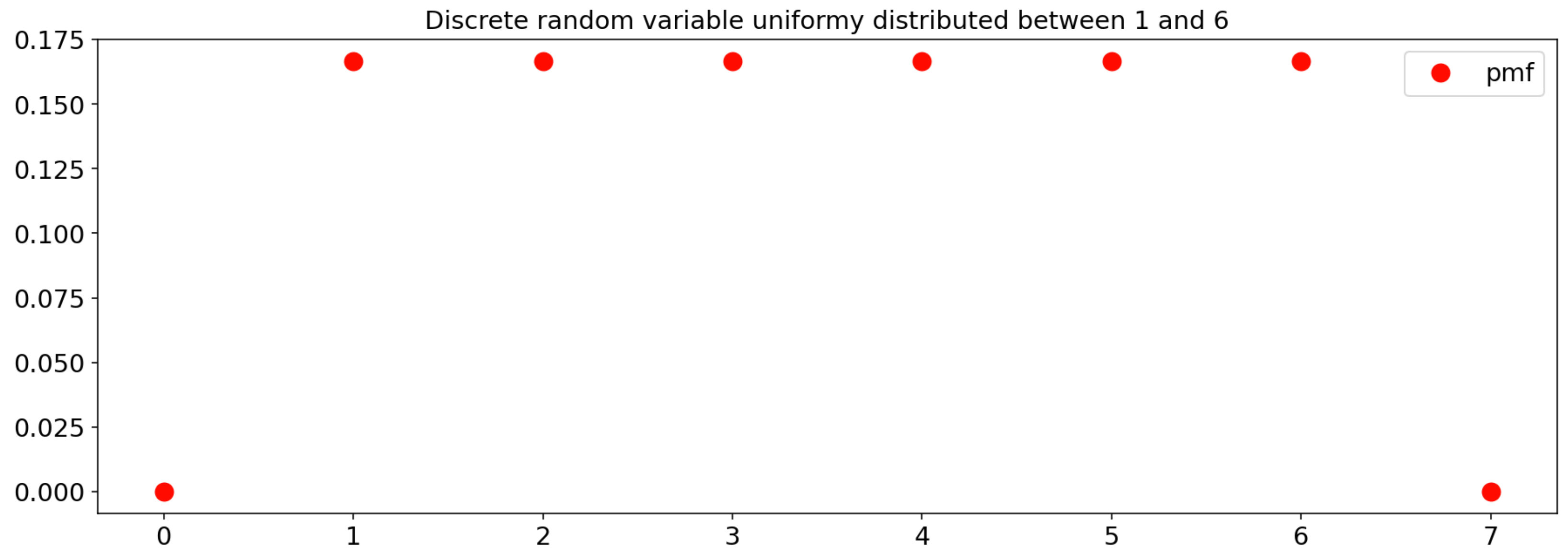
Process

↓

Distribution $p(x_i) = 1/6$

↓

Data



Statistics

Data

↓

Distribution

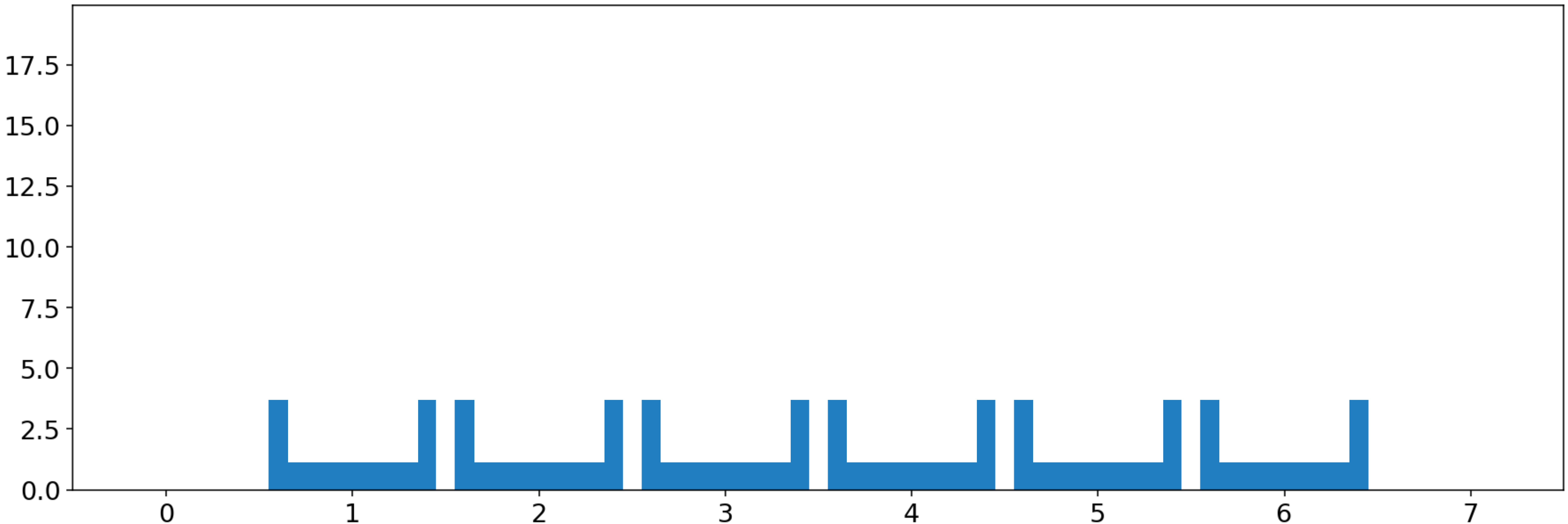
↓

Process

Histograms

An *approximate* representation of the distribution of a random variable

Definition: the frequency of realizations occurring in certain ranges of values (bins)

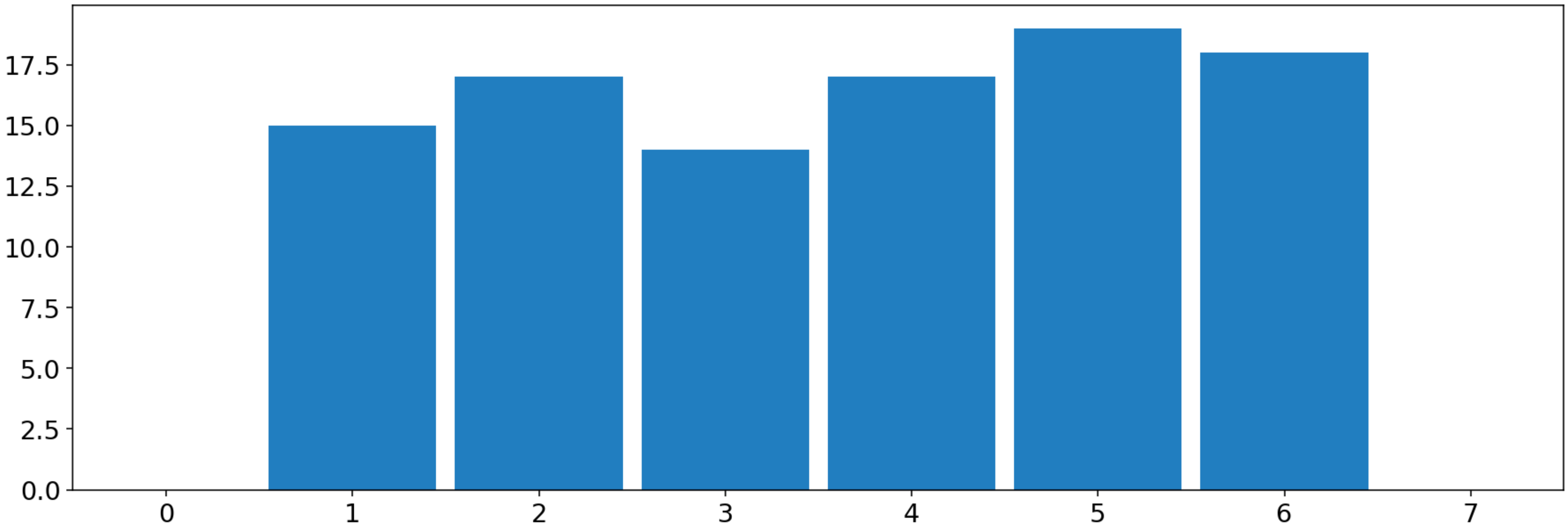


Histograms

An *approximate* representation of the distribution of a random variable

Definition: the frequency of realizations occurring in certain ranges of values (bins)

Count: the *number* of realizations occurring in each bin

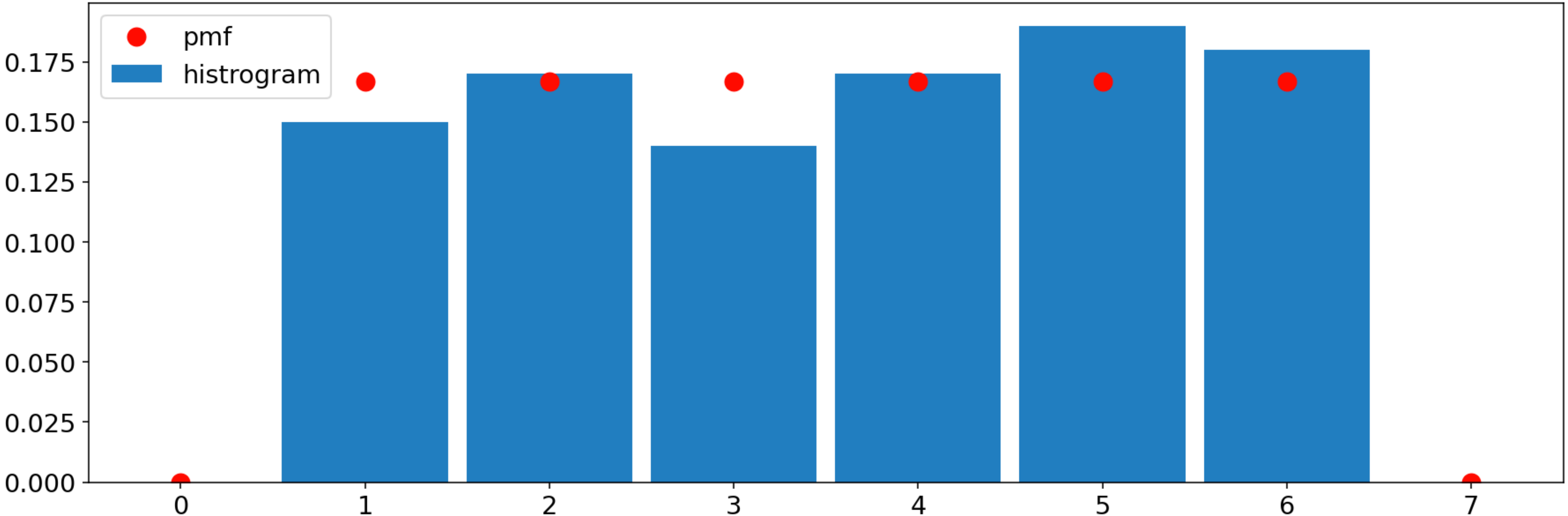


Histograms

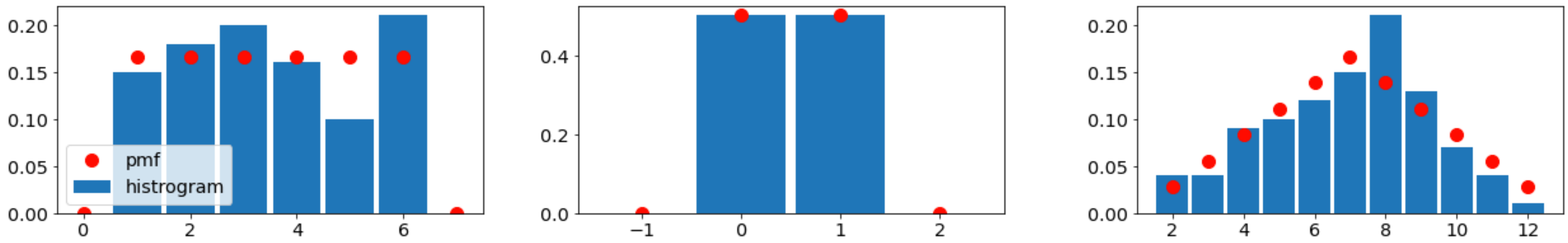
An *approximate* representation of the distribution of a random variable

Definition: the frequency of realizations occurring in certain ranges of values (bins)

Frequency: the *relative number* of realizations occurring in each bin
(Number in each bin / total number)

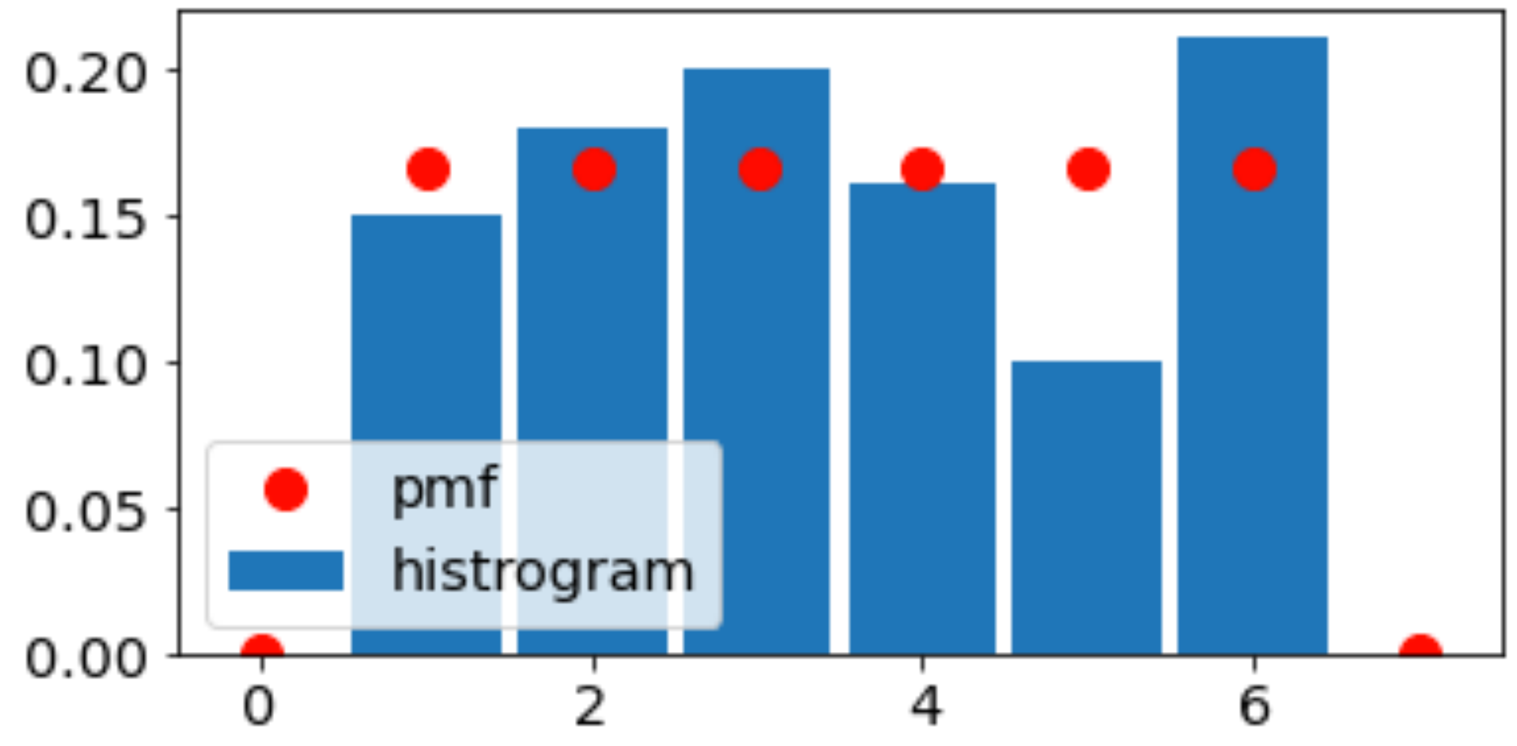


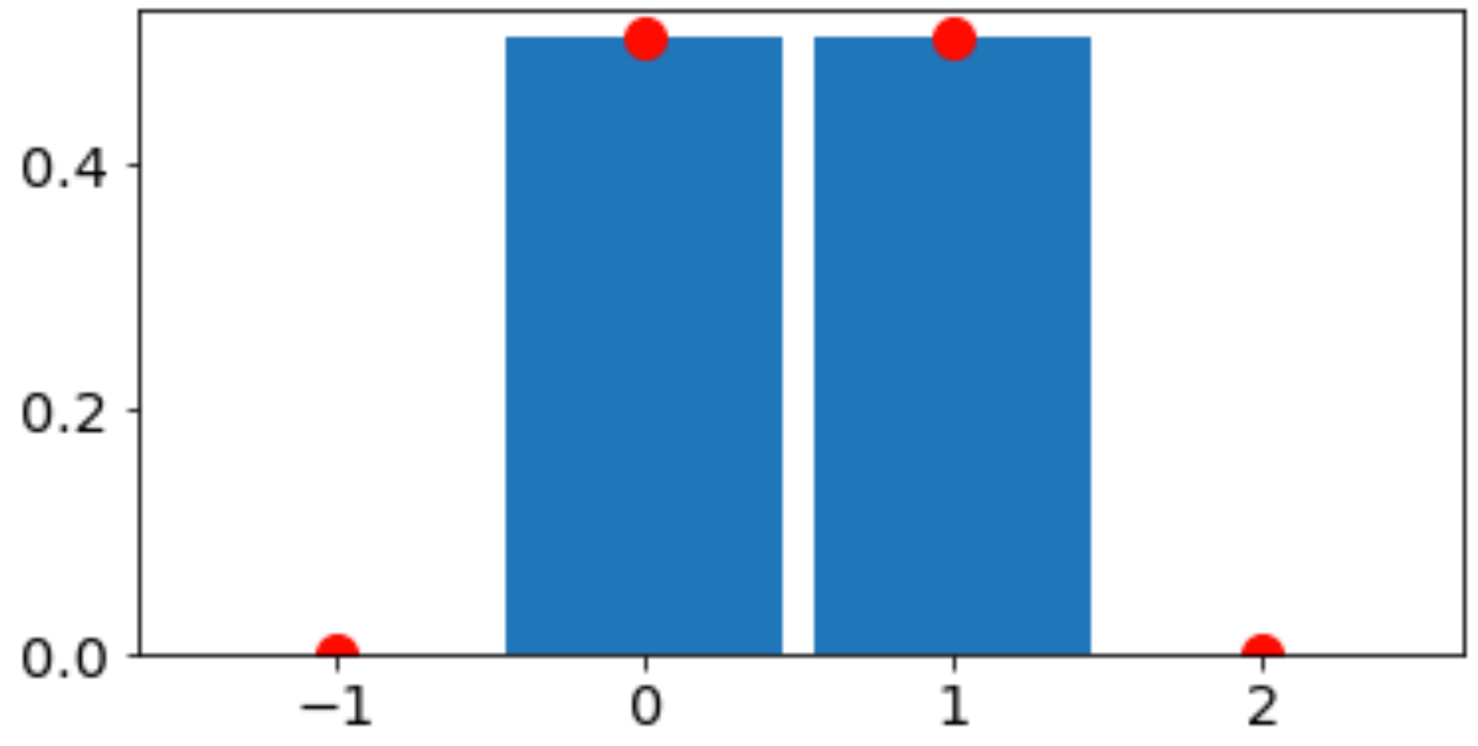
Histograms

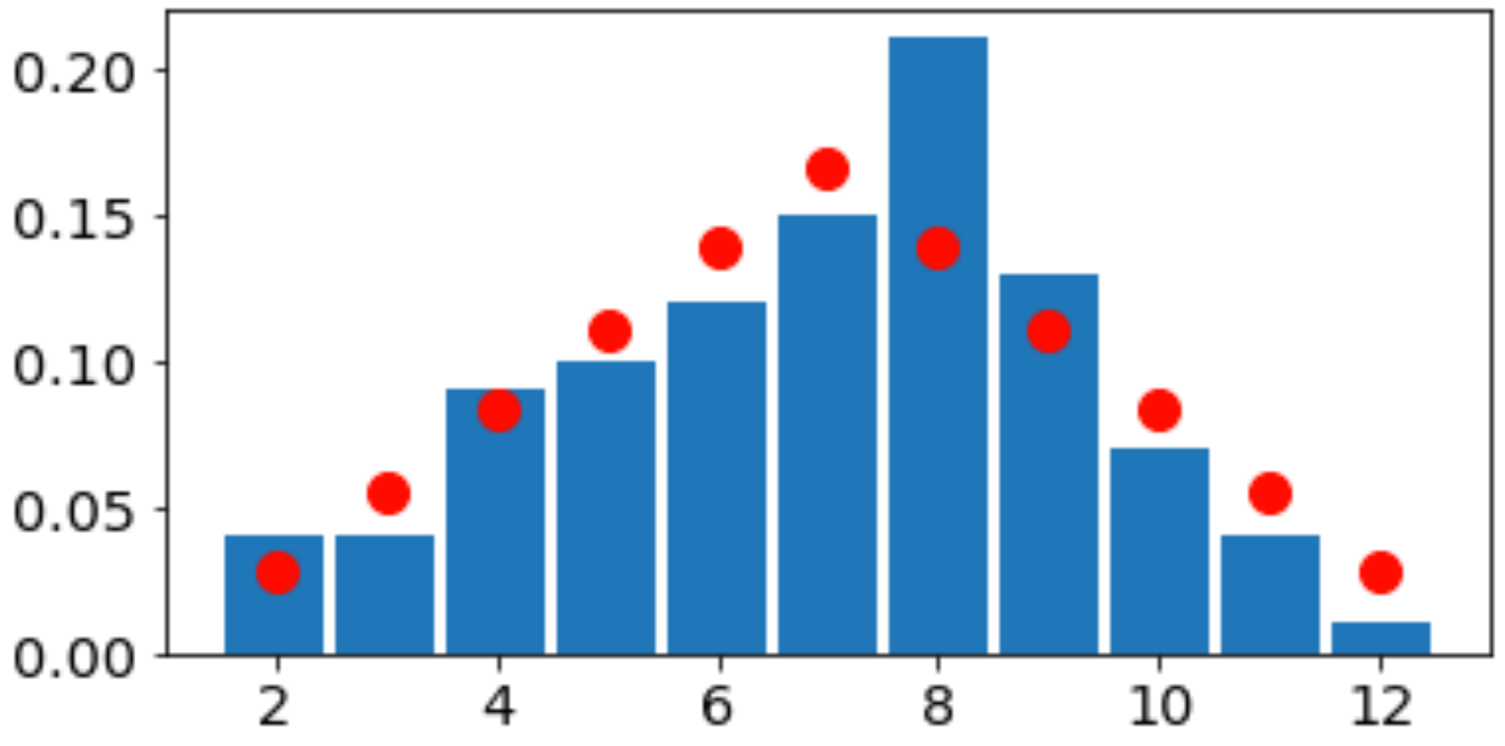


pmf
histrogram
0.00
0.05
0.10
0.15
0.20
0
2
4
6
0.0
0.2
0.4
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
2
4
6
8
10
12

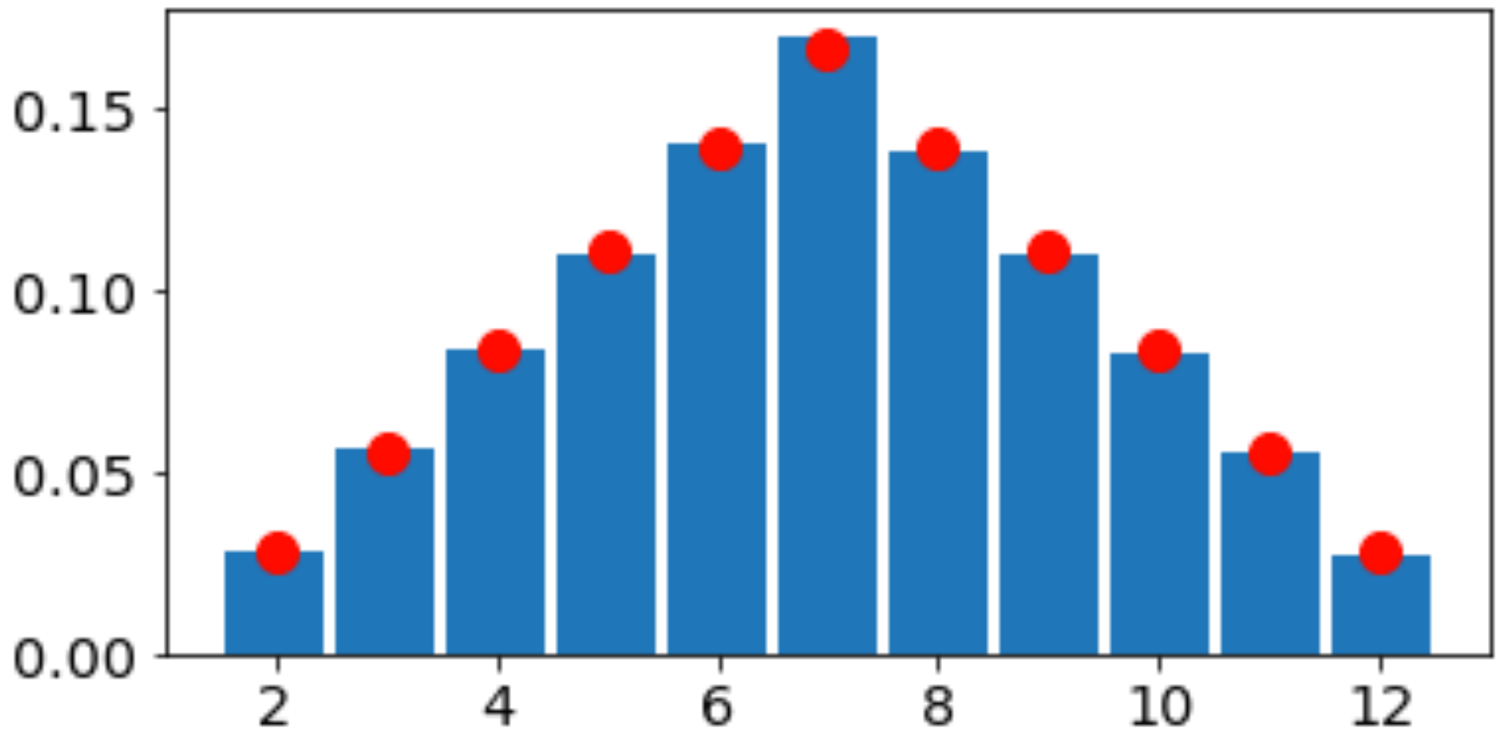
Histograms converge to pmfs







$$h_N(x_i) \to p(x_i)$$



Probability ↔ Statistics

Probability

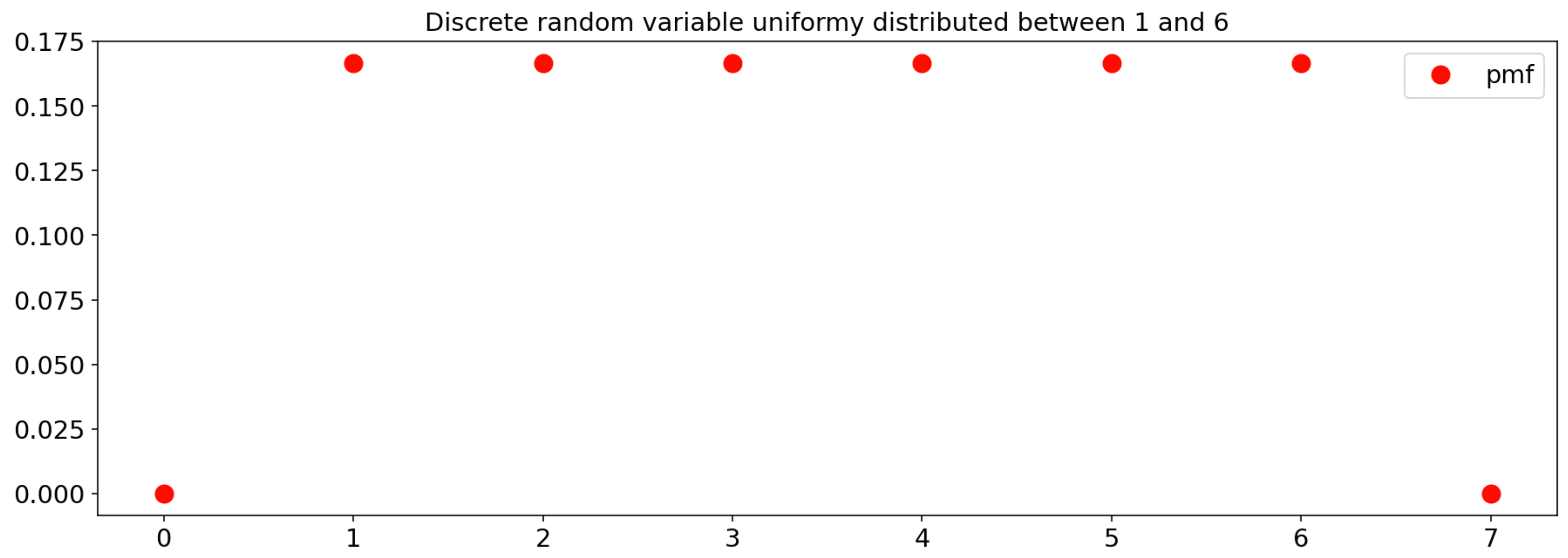
Process

↓

Distribution $p(x_i) = 1/6$

↓

Data



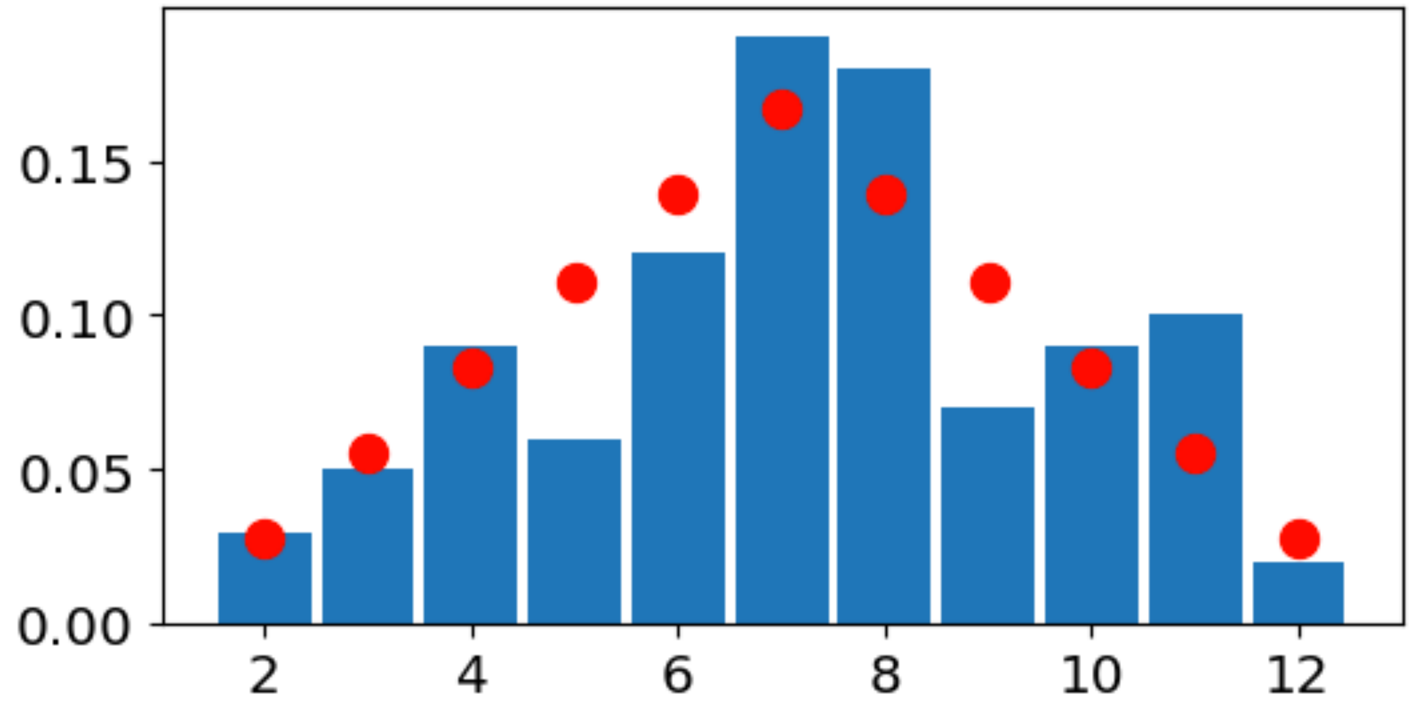
Statistics

Data

```
[11  4  4  2  6  7  8  5 11  2 10  8  8  8  9  6  8  4  3  7  7  8  5  7
  7 10  8  7  7  4  4  7  9  5 11 10 11  9 11  9  7  7  3  8  8  7 10  5
  6  4 11  8  9  5  3  8  4  6  8  2  7  6  3 10  9  4  8 10  7  8  6 10
  8  6 12 11  6  7  4  6 11  3  7 10 11  9  8  5  7  6 12  7  8 11  6 10
  6  8  7  7]
```

↓

Distribution



↓

Process

Continuous Random Variables

Reading:

- Emile-Geay: Chapter 3

Continous R.Vs
Moments of distribution
Sampling

Continuous Random Variables

A random variable X is a function defined on the sample space,
That associates a number for outcome and event:

Discrete:

$$X : \mathcal{S} \to \{x_i\}$$

Continuous:

$$X : \mathcal{S} \to \mathbb{R}$$

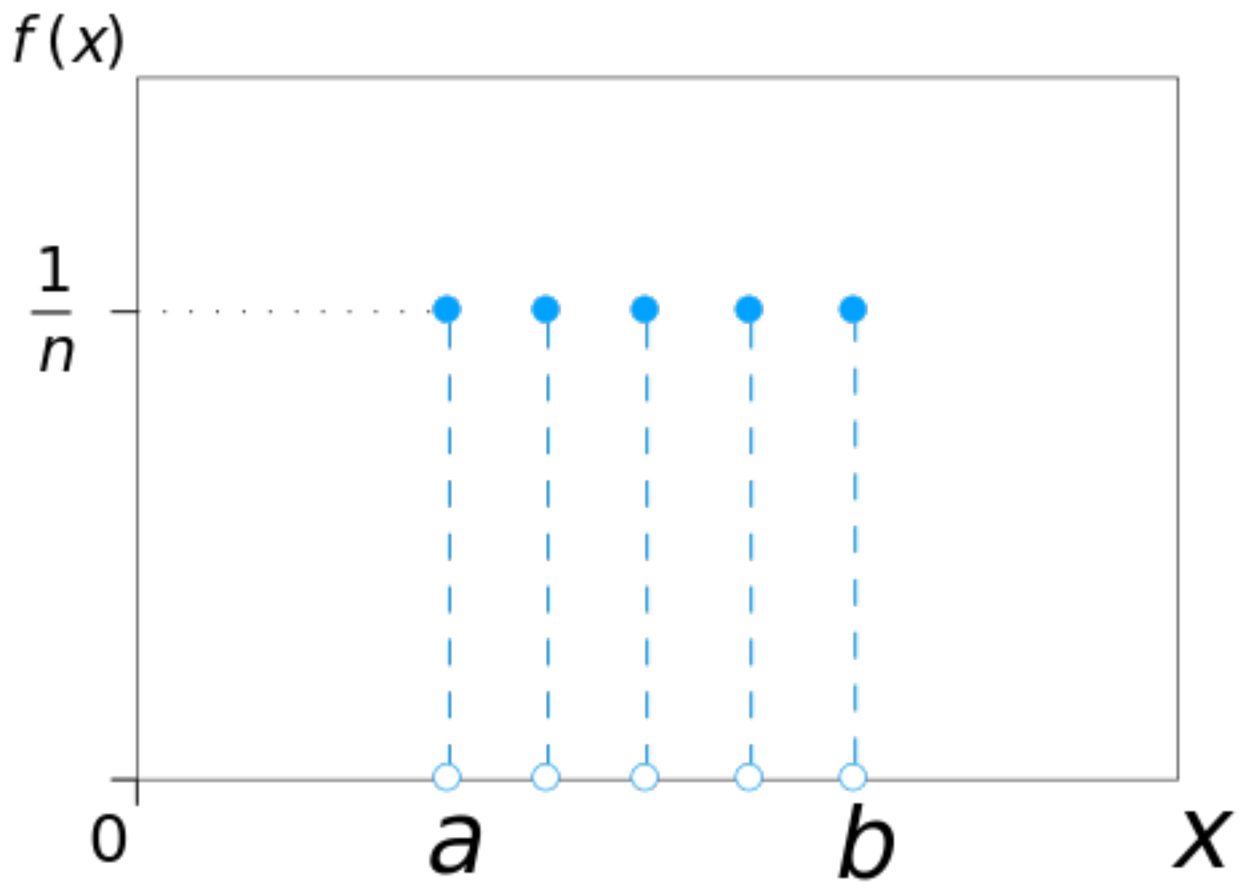
$$X : \mathcal{S} \to [a, b]$$

$$X : \mathcal{S} \to (a, b]$$

$$X : \mathcal{S} \to [0, \infty)$$

Distributions: Cumulative Distribution Function

X: uniform discrete distribution over $x_i, i \in \{1,n\}$

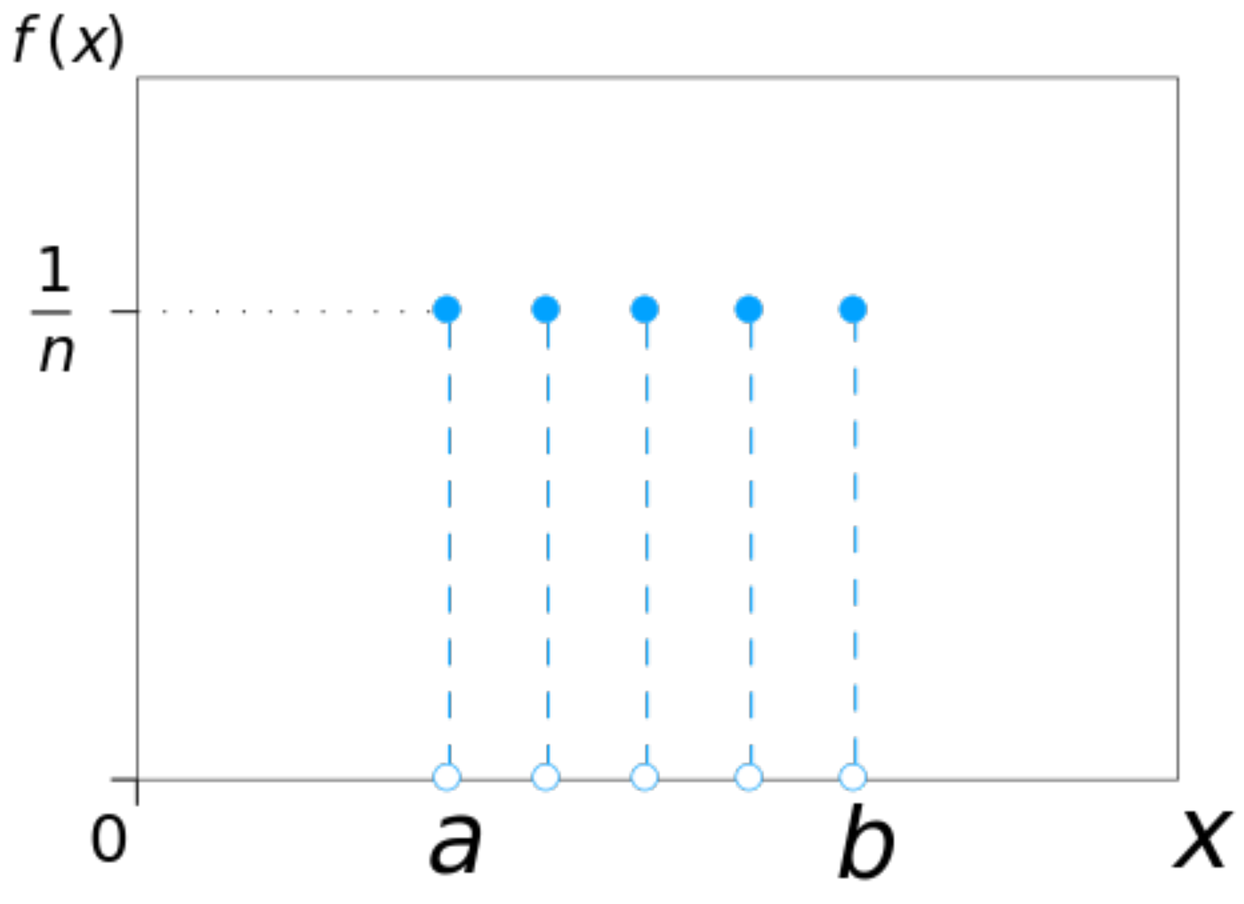


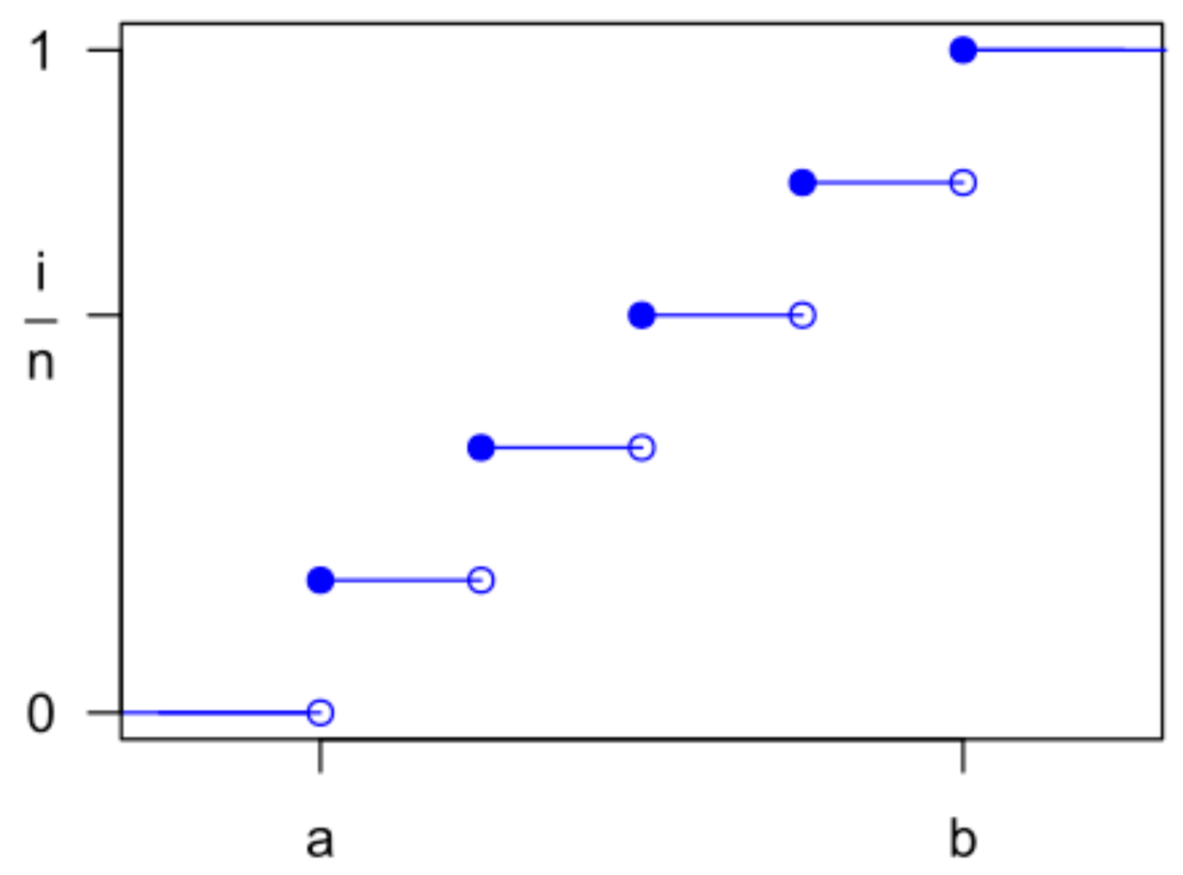
Probability Mass Function

$$f(x_i) = P(X = x_i)$$

Distributions: Cumulative Distribution Function

X: uniform discrete distribution over $x_i, i \in \{1,n\}$





Probability Mass Function

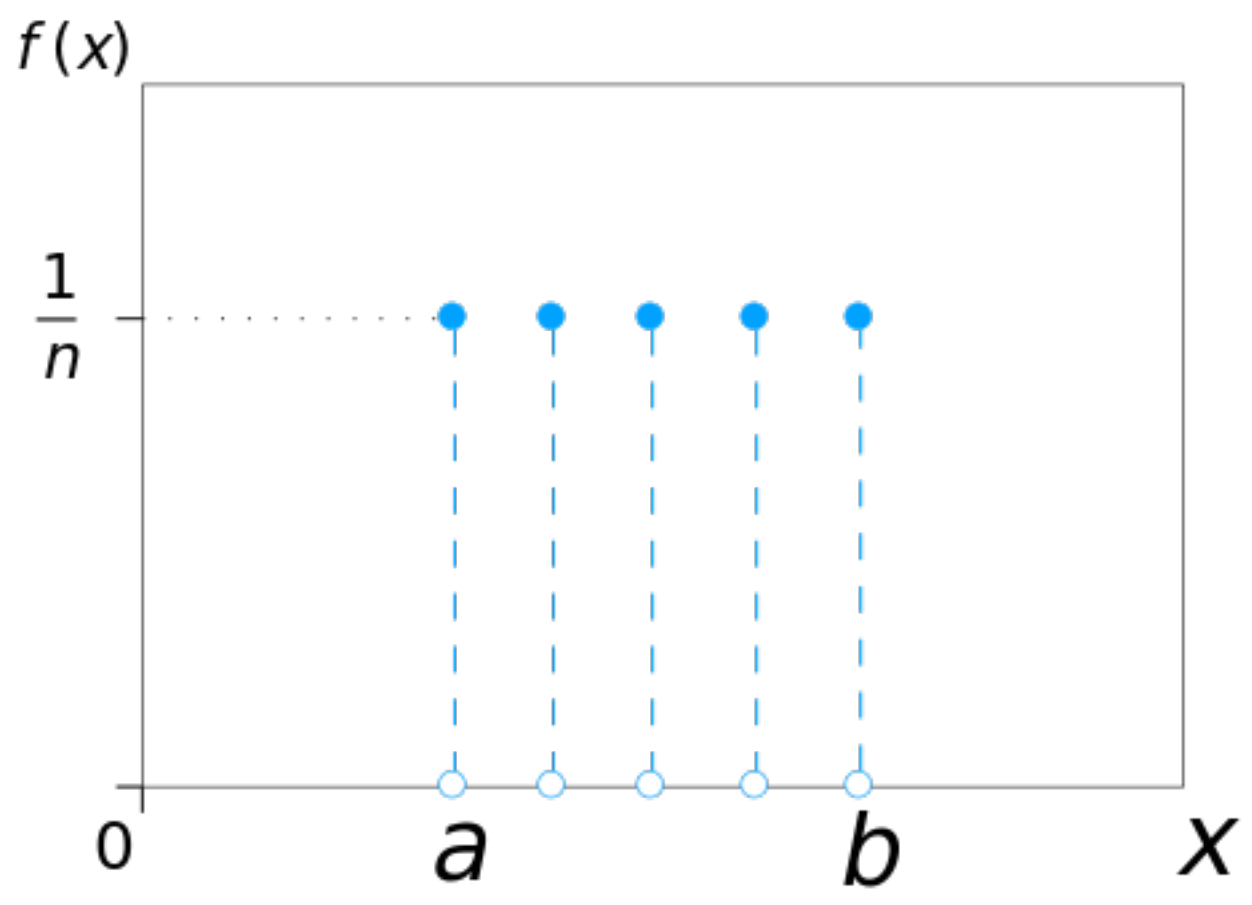
$$f(x_i) = P(X = x_i)$$

Cumulative Distribution Function

$$F(x) = P(X \leq x)$$

Distributions: Cumulative Distribution Function

X: uniform discrete distribution over $x_i, i \in \{1,n\}$

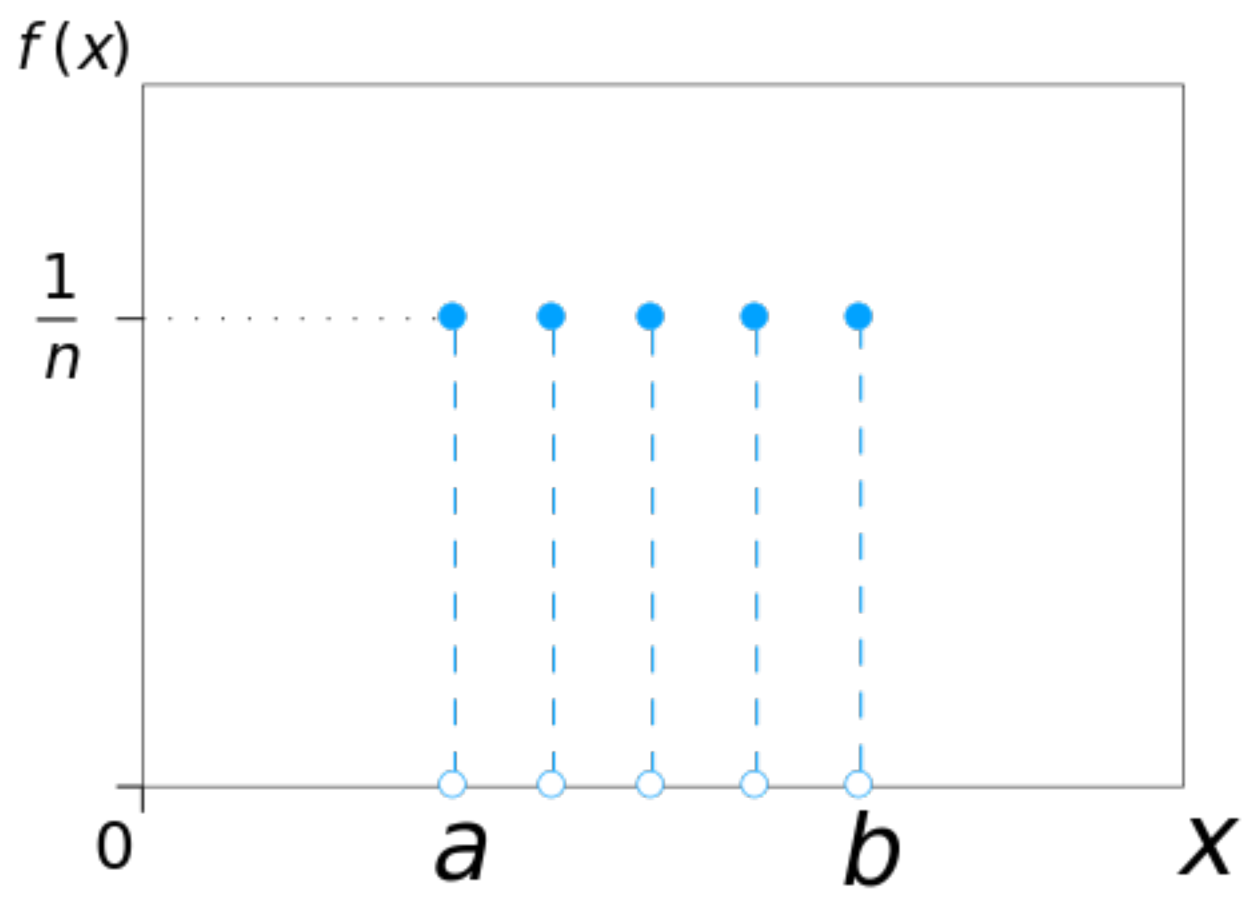


PMF

$$f(x_i) = P(X = x_i)$$

Distributions: Cumulative Distribution Function

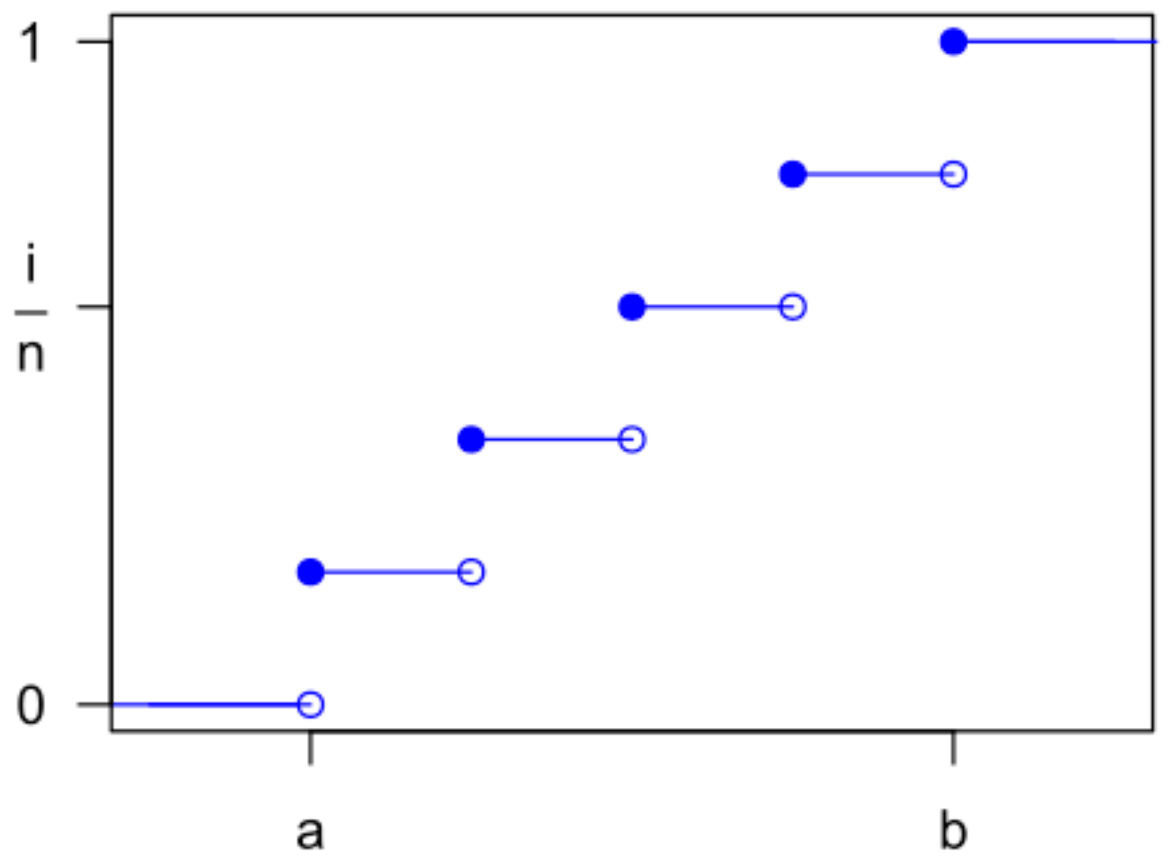
X: uniform discrete distribution over $x_i, i \in \{1, n\}$



PMF

$$f(x_i) = P(X = x_i)$$

$$F(x) = \sum_{i:x_i<x} f(x_i)$$



CDF

$$F(x) = P(X \leq x)$$

Distributions: Probability Density Function (pdf)

pmf:

$$f(x_i) = P(X = x_i)$$

Distributions: Probability Density Function (pdf)

pmf:

$$f(x_i) = P(X = x_i)$$

pdf:

$$f(x) = P(X \in [x, x + dx])$$

$$P(x \in [a, b]) = \int_a^b f(x)dx$$

Distributions: Probability Density Function (pdf)

pmf:

$$f(x_i) = P(X = x_i)$$

pdf:

$$f(x) = P(X \in [x, x + dx])$$

$$P(x \in [a, b]) = \int_a^b f(x)dx$$

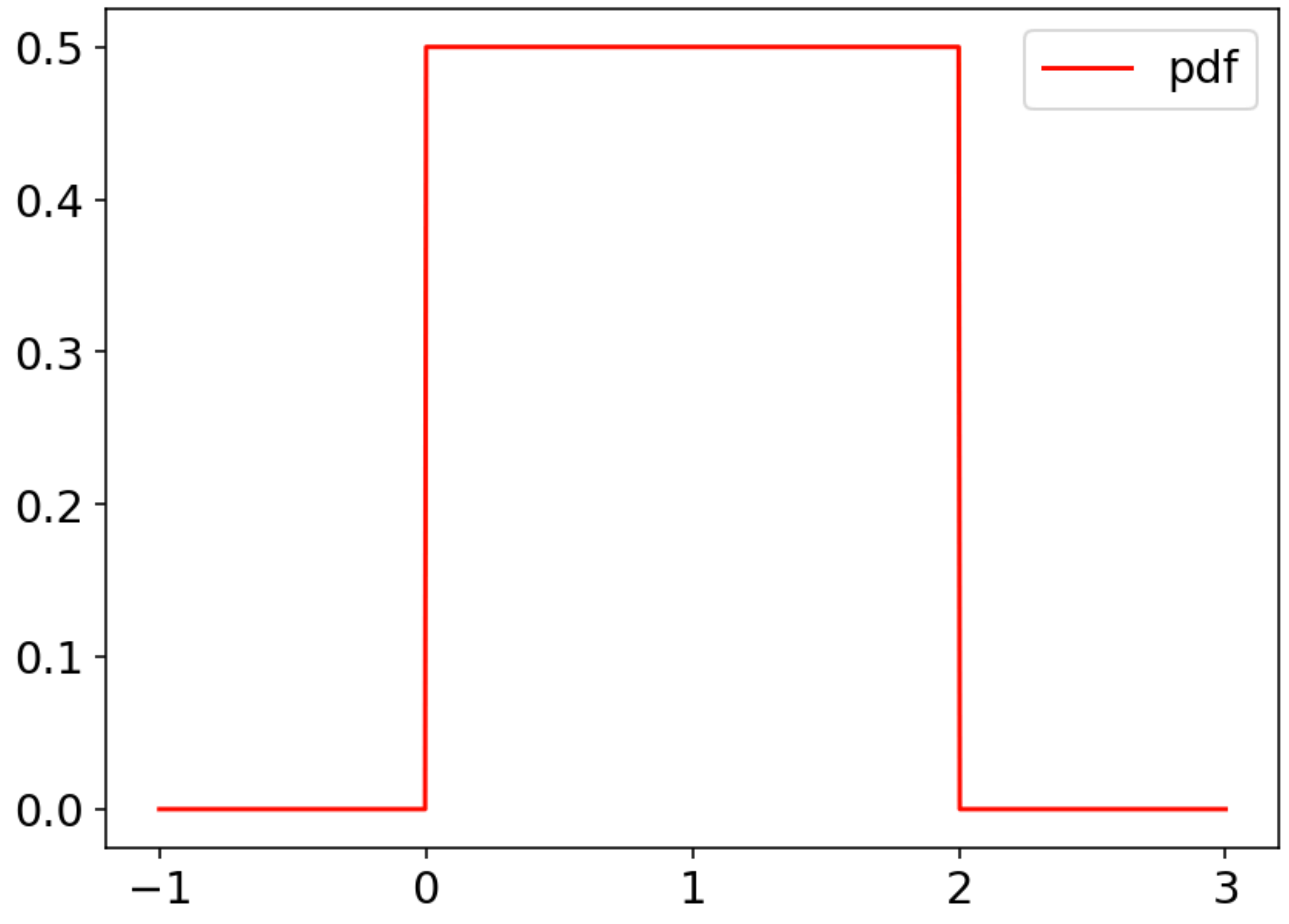
cdf

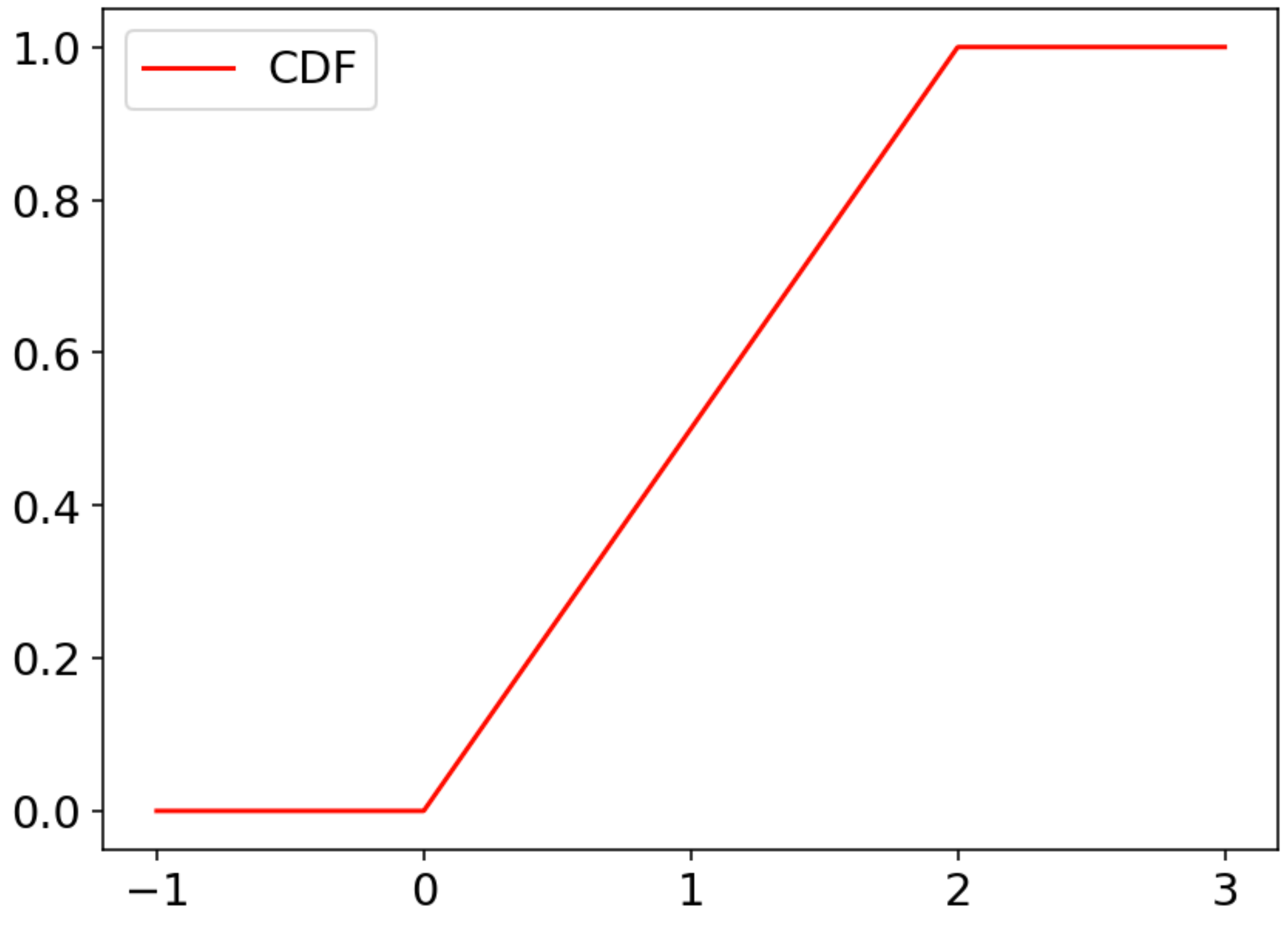
$$F(x) = P(X < x) = \int_{-\infty}^{x} f(u)du$$

Continuous uniform random variable

The uniform distribution is the distribution of a random variable which has equal probabiliy of taking any value within an interval a,b

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$$



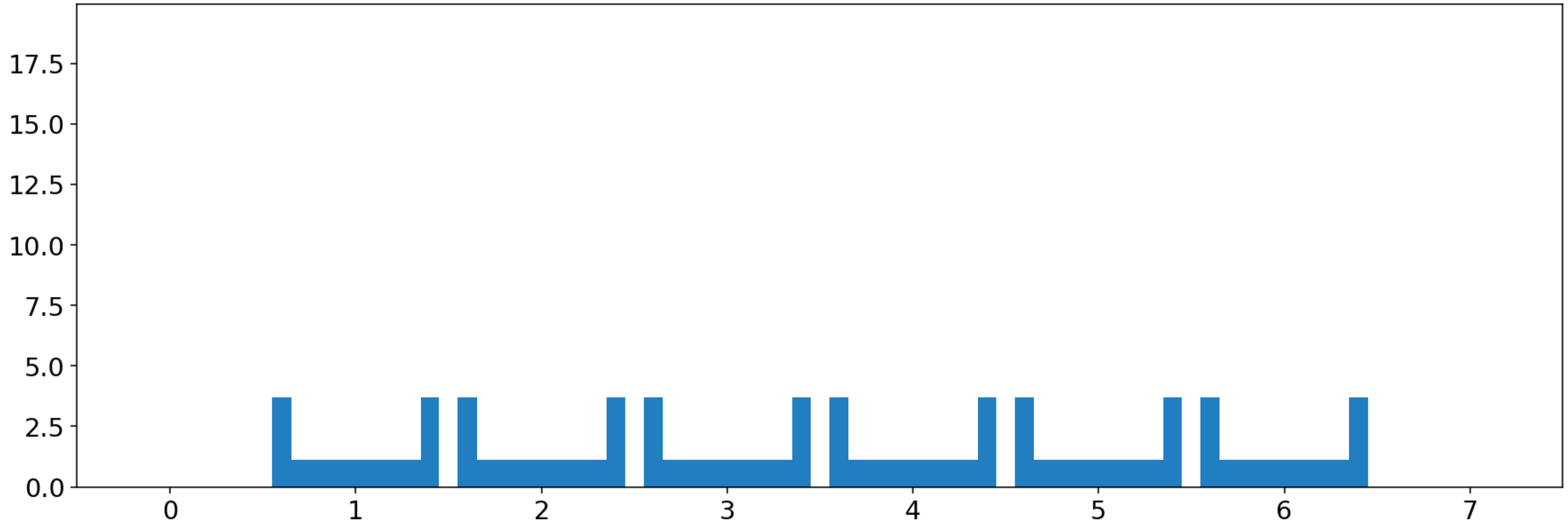


Histograms

Definition: the frequency of realizations occurring in certain ranges of values (bins)

$$F(x) = P(X < x) = \int_{-\infty}^{u} f(u)du$$

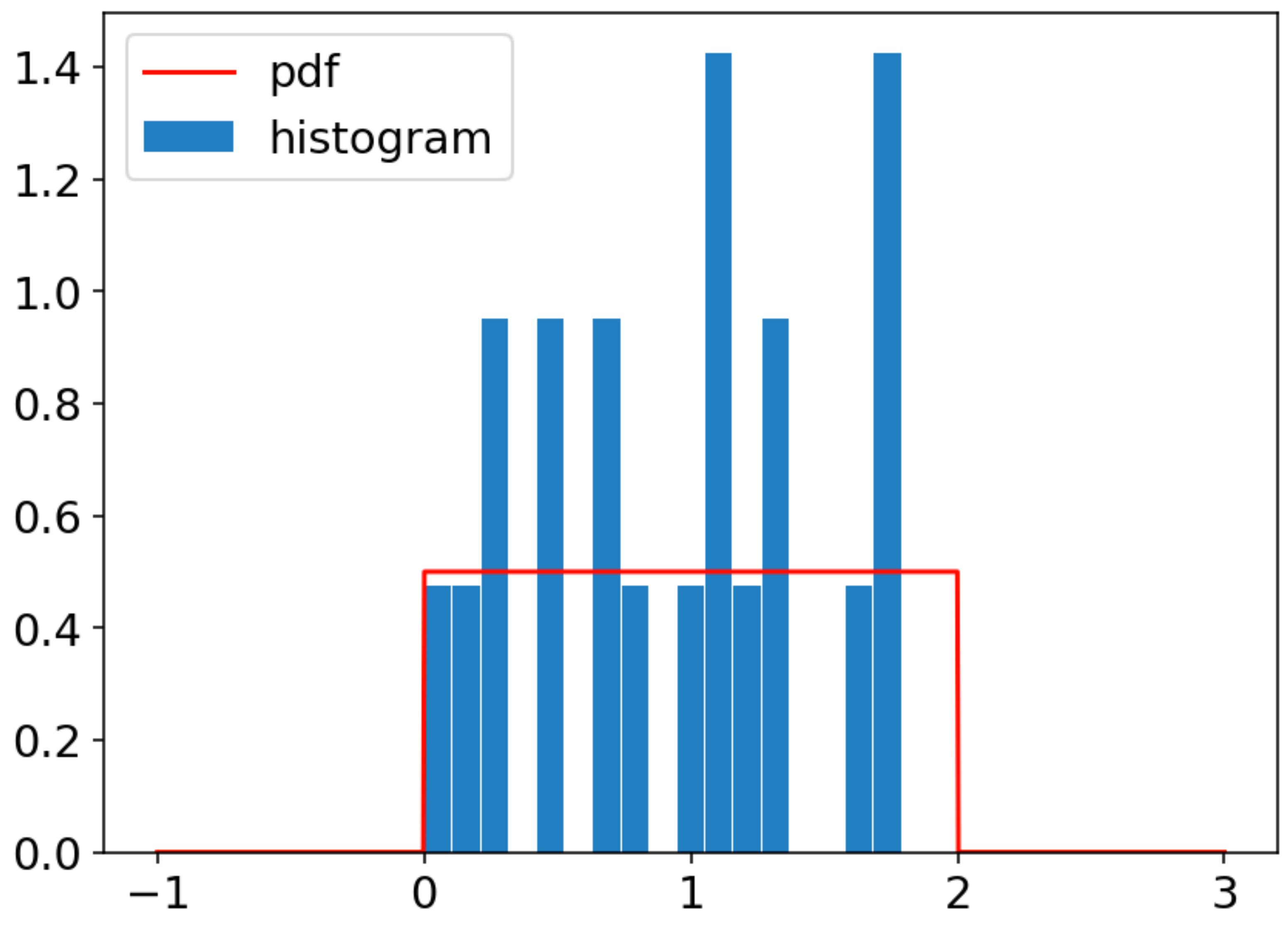
$$F(x) = P(X < x) = \sum_{i:x_i<x} f(x_i)$$

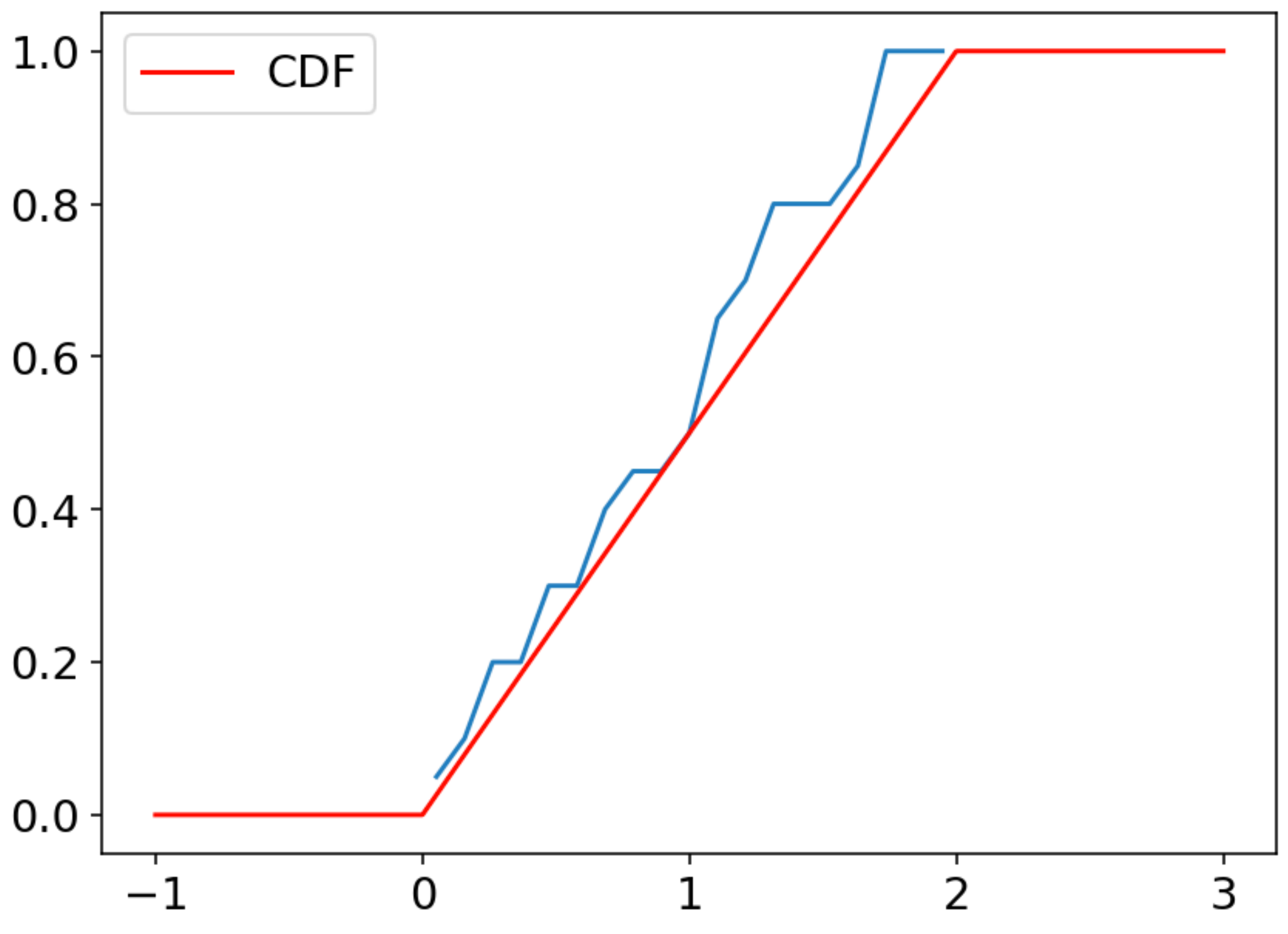


Continuous uniform random variable

The uniform distribution is the distribution of a random variable which has equal probabiliy of taking any value within an interval a,b

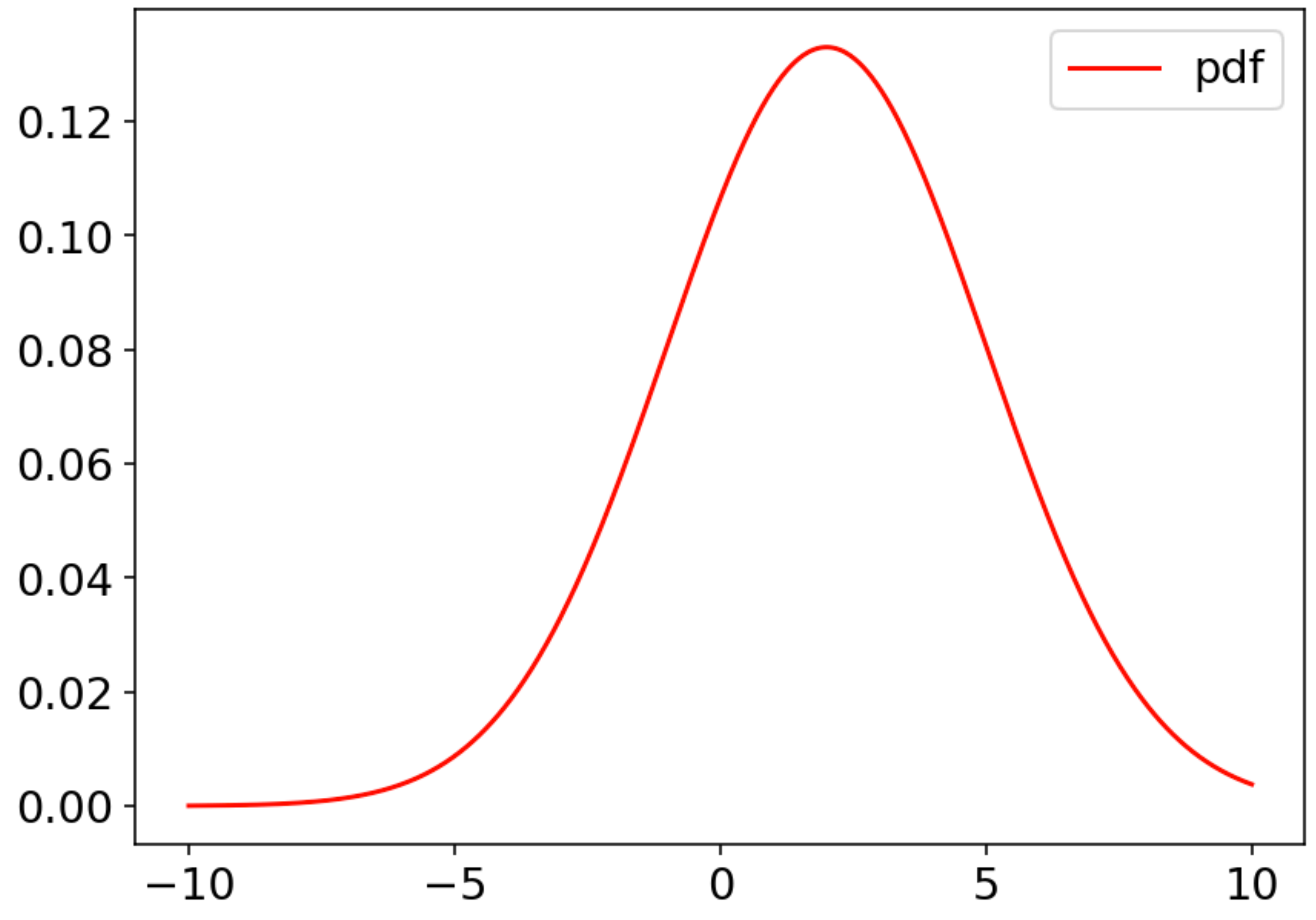
$$p(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$$

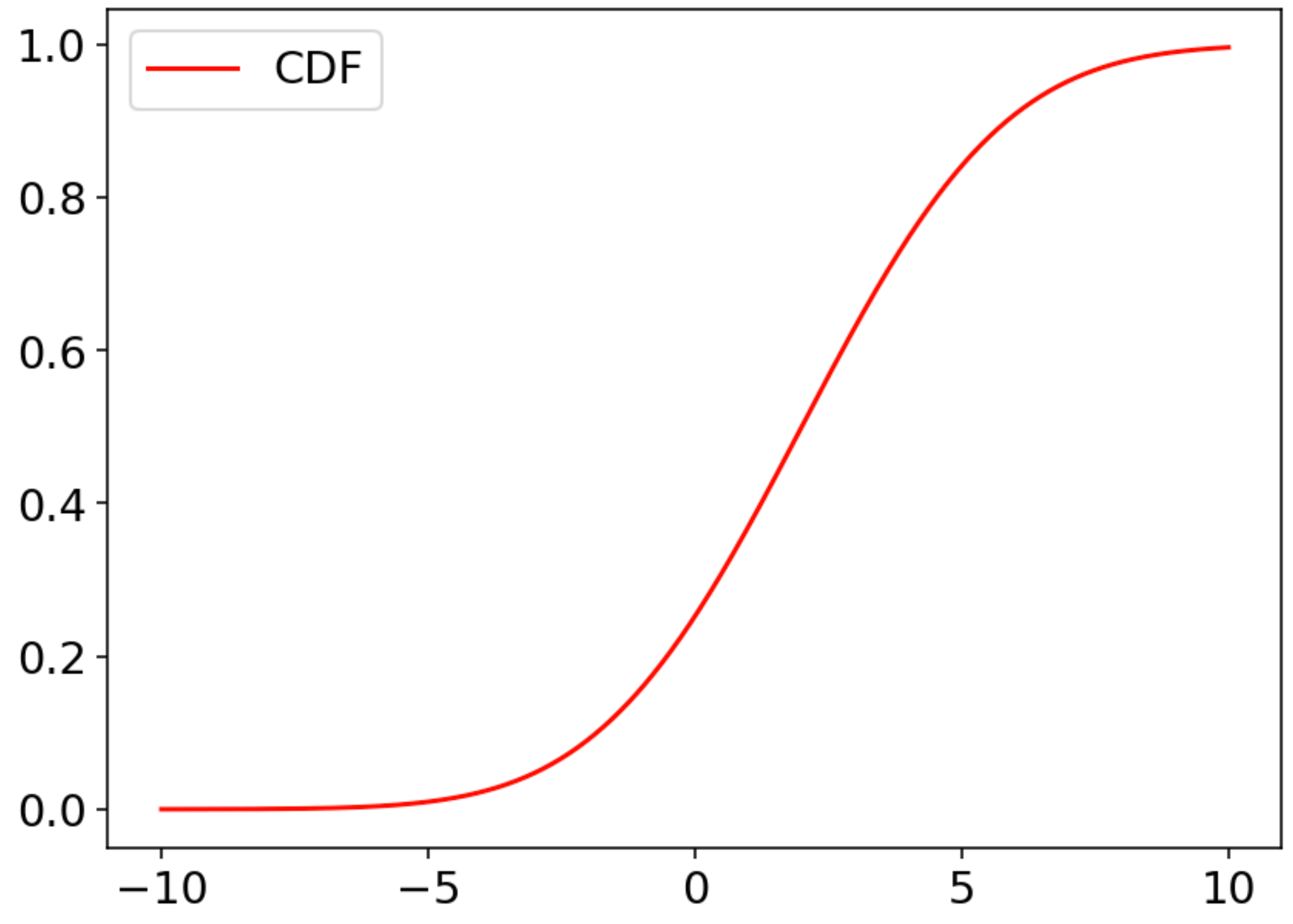




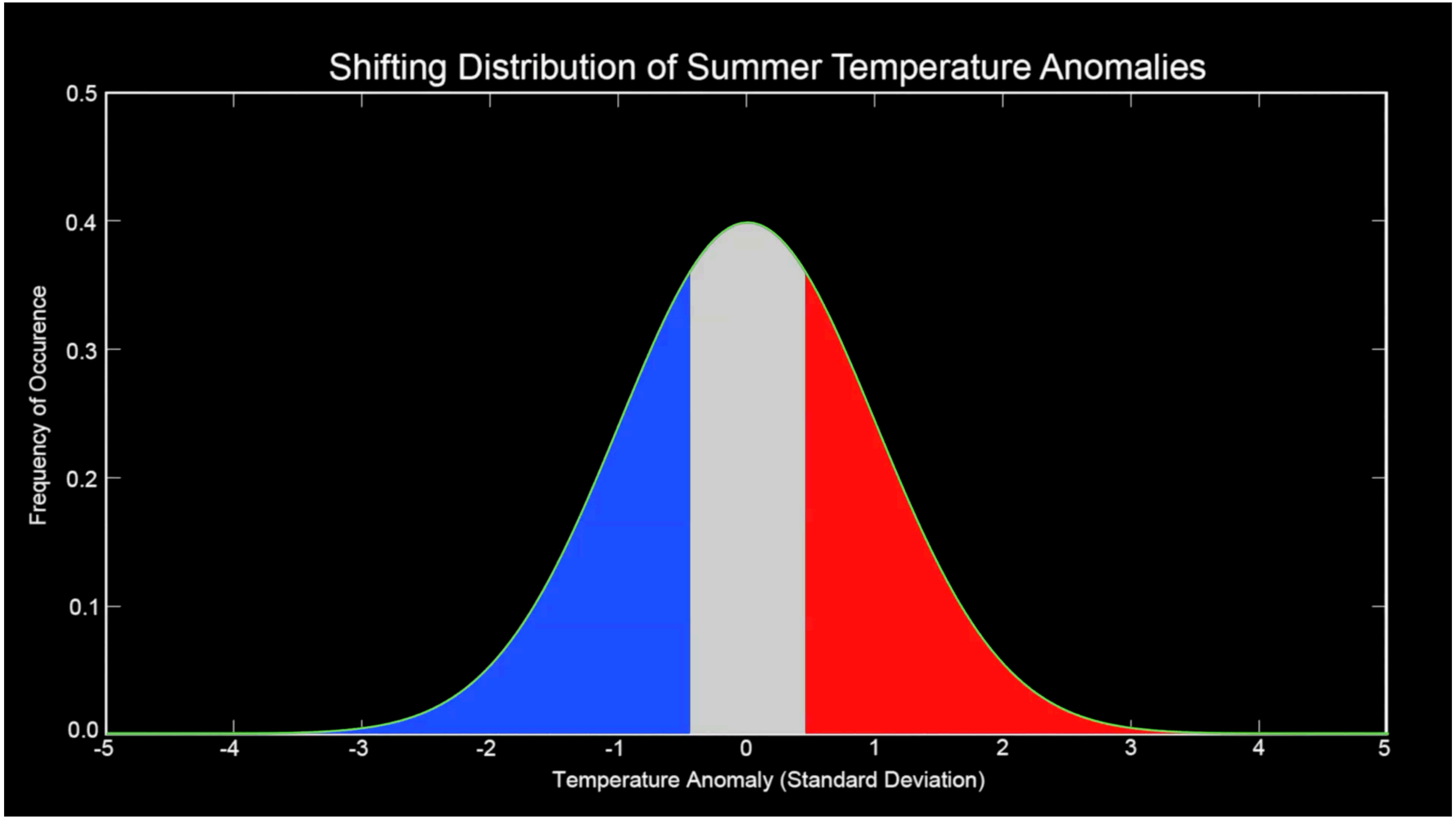
Normal/Gaussian Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

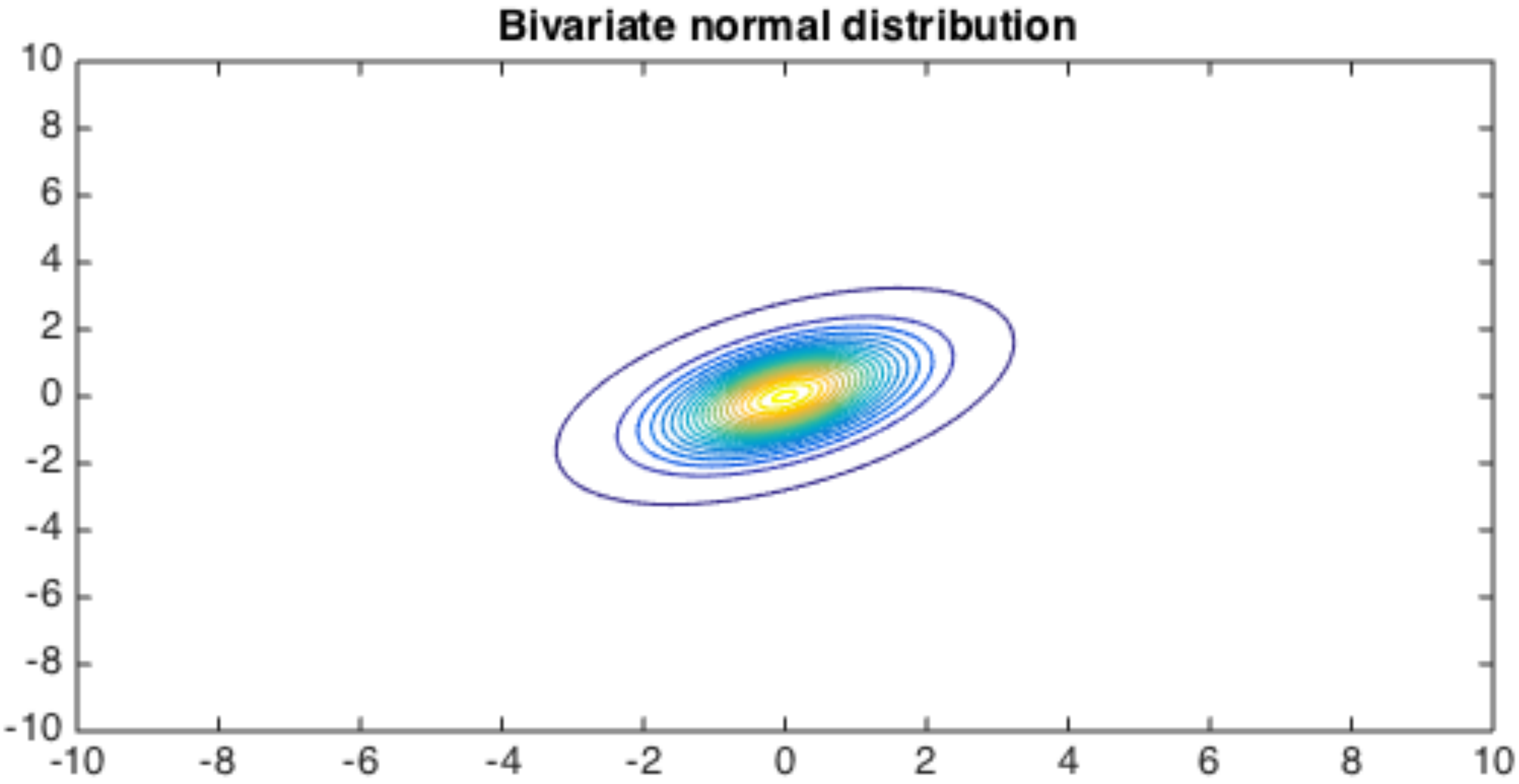




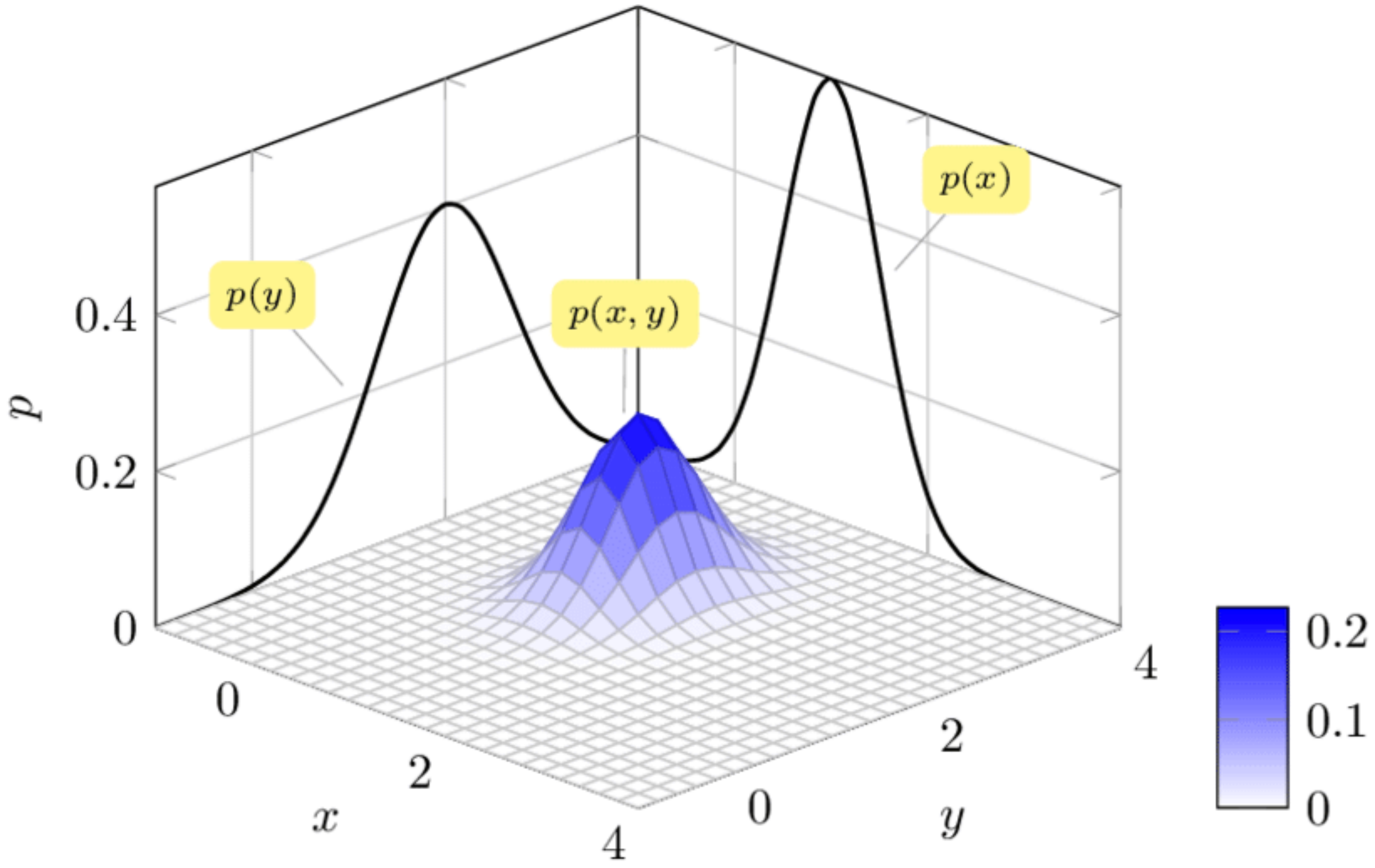
Real World distributions



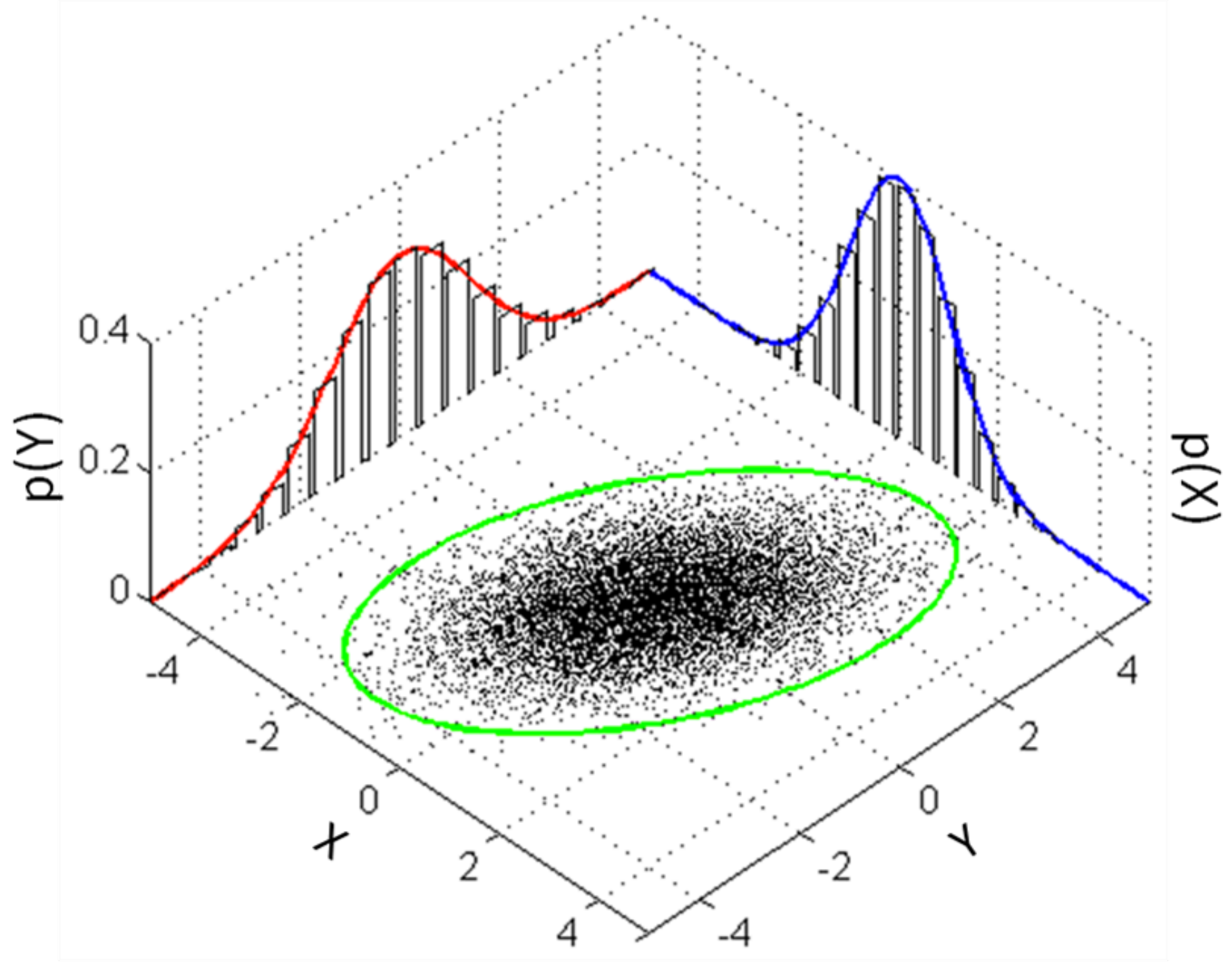
Multivariate Distributions



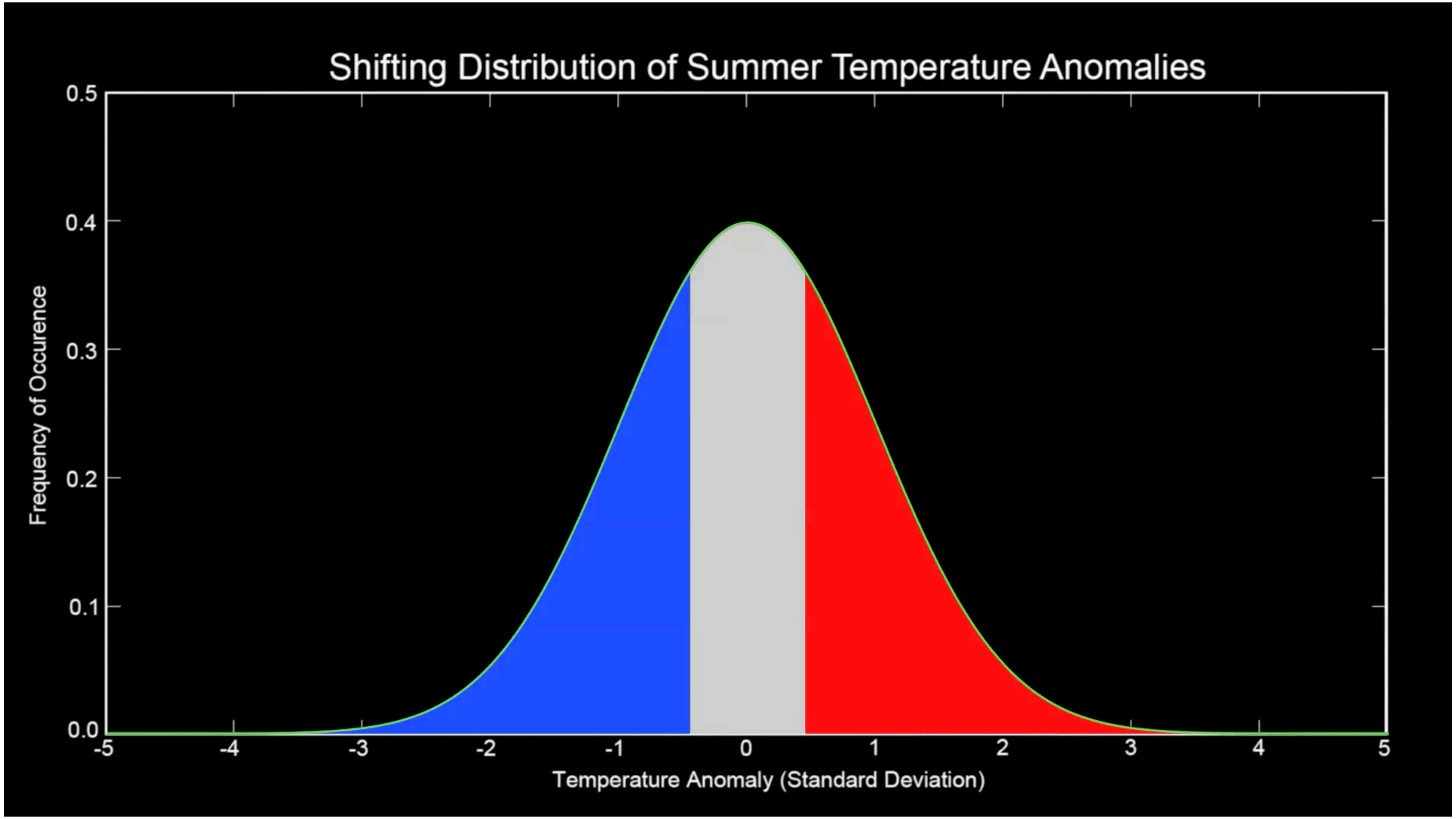
Bivariate Gaussian with marginals: distribution



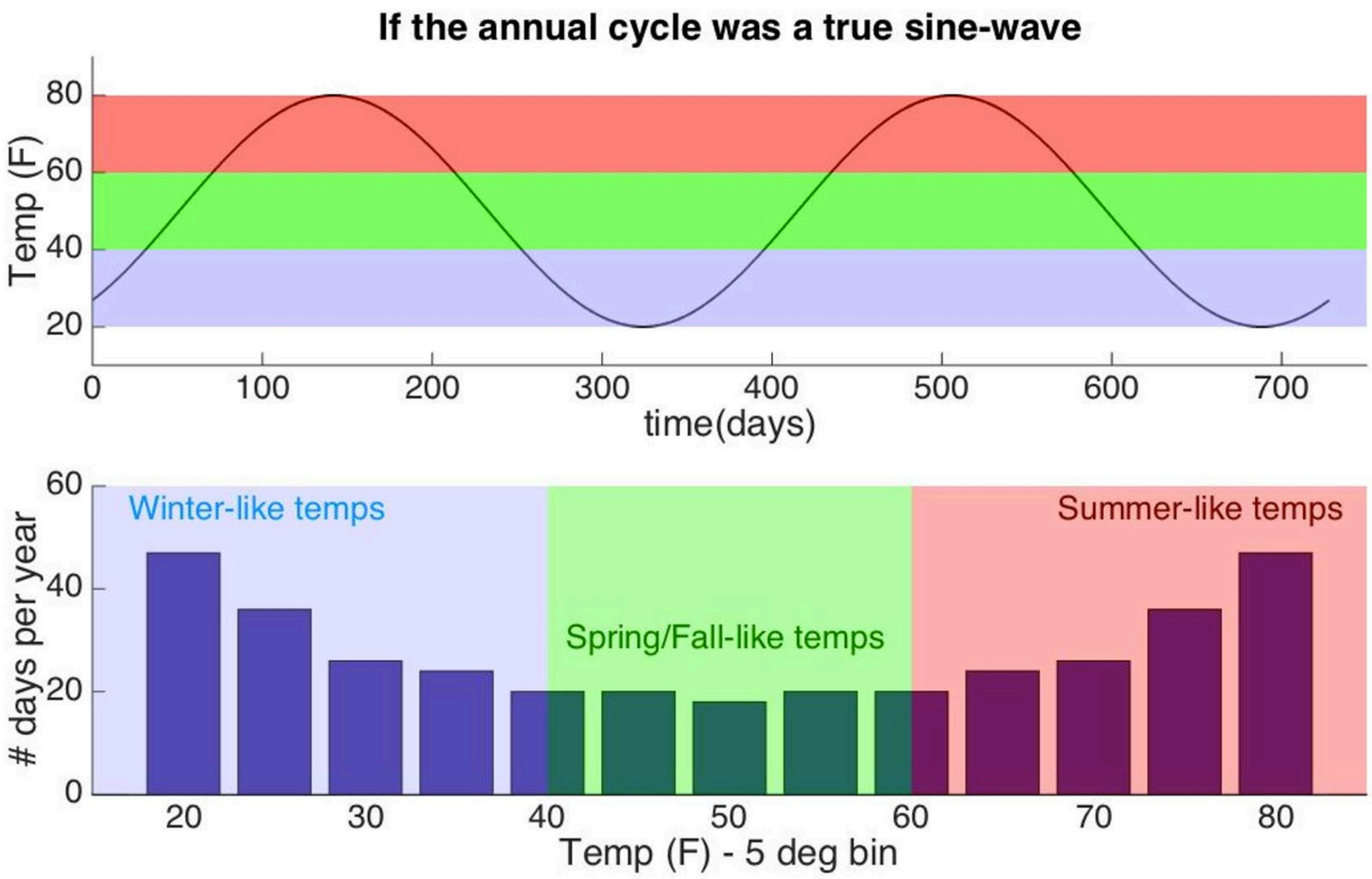
Bivariate Gaussian with marginals: histogram



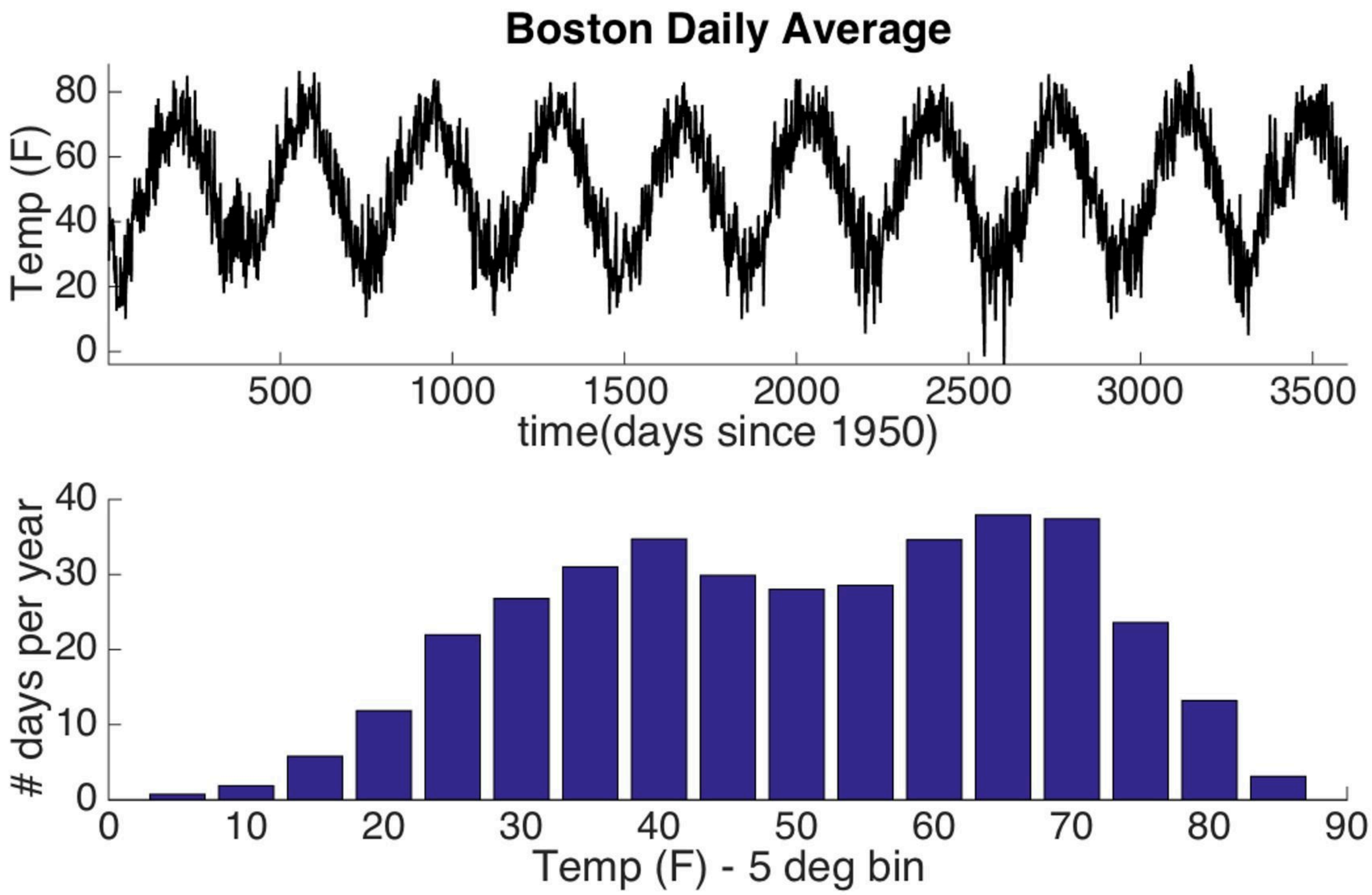
Real World distributions



Mixed Distributions



Mixed Distributions



Mixed Distributions