# Probability Theory

CMPUT 267: Basics of Machine Learning

§2.1-2.2

# Recap for the Course Start

> This class is about **understanding** machine learning techniques by understanding their basic **mathematical underpinnings**

- Please read the FAQ and Getting Started (it will save us all time)
- Assignment 1 released
- Thought Questions 1 due very soon (January 20)
  - Biggest reading since it covers much of the background
- Updates to notes throughout, to fix a few typos and add clarity (redownload)

# Outline

1. Probabilities
2. Defining Distributions
3. Random Variables

# Why Probabilities?

Even if the world is completely deterministic, outcomes can **look random**

**Example:** A high-tech gumball machine behaves according to
$f(x_1, x_2) =$ output candy if $x_1$ & $x_2$,
where $x_1$ = **has candy** and $x_2$ = **battery charged**.

- You can only see if it has candy (only see $x_1$)
- From your perspective, when $x_1 = 1$, sometimes candy is output, sometimes it isn't
- It **looks stochastic**, because it depends on the hidden input $x_2$

# Measuring Uncertainty

- **Probability** is a way of **measuring** uncertainty
- We assign a number between 0 and 1 to **events** (hypotheses):
  - **0** means absolutely certain that statement is **false**
  - **1** means absolutely certain that statement is **true**
  - **Intermediate** values mean more or less certain
- Probability is a measurement of **uncertainty**, **not truth**
  - A statement with probability .75 is not "mostly true"
  - Rather, we **believe** it is more **likely** to be true than not

# Example

- Let's think about estimating the average height of a person in the world
- There is a true population mean $h$ (say $h$ = 165.2 cm)
  - which can be computed by averaging the heights of every person
- We can estimate this true mean using data
  - e.g., compute a sample average $\bar{h}$ from a subpopulation by randomly sampling 1000 people from around the whole world (say $\bar{h}$ = 166.3 cm)
- We can also reason about our belief over plausible estimates $\bar{h}$ of $h$
  - e.g., we can maintain a distribution over plausible $\bar{h}$, such as saying $p(\bar{h} = 160) = 0.1$, $p(\bar{h} = 163) = 0.3$, $p(\bar{h} = 165) = 0.5$, $p(\bar{h} = 167) = 0.1$

# Terminology Refresher

- Chapter 1 has a refresher and some exercises, and there is a notation sheet at the beginning of the notes
- Set notation
  - Curly brackets for discrete sets, e.g $\{a, b, c\}$, $\{1,2,3,4,5\}$, $\{-2.1,6.5\}$
  - Square brackets for continuous intervals, e.g., $[-10,10]$, $[3.2,7.1]$
  - Subset notation $A \subset \Omega$ and the set complement $A^c = \Omega \backslash A$
  - Union of sets $A \cup B$, intersection of sets $A \cap B$
  - Power set $\mathscr{P}(A)$, e.g, $A = \{1,2\}$, $\mathscr{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$
- Scalar $x \in \mathbb{R}$ and vector (array) is $\mathbf{x} \in \mathbb{R}^d$ for some integer $d \in \{2,3,\ldots\}$

# Terminology (cont.)

- **Countable:** A set whose elements can be assigned an integer index
  - The integers themselves
  - Any finite set, e.g., $\{0.1, 2.0, 3.7, 4.123\}$
  - We'll sometimes say **discrete**, even though that's a little imprecise
- **Uncountable:** Sets whose elements *cannot* be assigned an integer index
  - Real numbers $\mathbb{R}$
  - Intervals of real numbers, e.g., $[0,1]$, $(-\infty, 0)$
  - Usually we'll say we have a **continuous set**

# Outcomes and Events

All probabilities are defined with respect to a **measurable space** $(\Omega, \mathscr{E})$ of **outcomes** and **events**:

- $\Omega$ is the **sample space**: The set of all possible outcomes
- $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ is the **event space**: A set of subsets of $\Omega$ that satisfies two key properties (that I will define in two slides)

# Examples of Discrete & Continuous Sample Spaces and Events

**Discrete (countable) outcomes**

$\Omega = \{1,2,3,4,5,6\}$

$\Omega = \{\text{person}, \text{robot}, \text{camera}, \text{TV}, \ldots\}$

$\Omega = \mathbb{N}$

**Continuous (uncountable) outcomes**

$\Omega = [0,1]$

$\Omega = \mathbb{R}$

$\Omega = \mathbb{R}^k$

# Event Spaces

**Definition:**

A non-empty set $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ is an **event space** if it satisfies

1. $A \in \mathscr{E} \implies A^c \in \mathscr{E}$
2. $A_1, A_2, \ldots \in \mathscr{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathscr{E}$

1. A collection of outcomes (e.g., either a 2 or a 6 were rolled) is an event.
2. If we can measure that an event has occurred, then we should also be able to measure that the event has not occurred; i.e., its **complement** is measurable.
3. If we can measure two events separately, then we should be able to tell if one of them has happened; i.e., their **union** should be measurable too.

# Examples of Discrete & Continuous Sample Spaces and Events

**Discrete (countable) outcomes**

$\Omega = \{1,2,3,4,5,6\}$

$\Omega = \{\text{person}, \text{robot}, \text{camera}, \text{TV}, \ldots\}$

$\Omega = \mathbb{N}$

$\mathscr{E} = \{\emptyset, \{1,2\}, \{3,4,5,6\}, \{1,2,3,4,5,6\}\}$

Typically**:** $\mathscr{E} = \mathscr{P}(\Omega)$

**Powerset** is the set of all subsets

**Continuous (uncountable) outcomes**

$\Omega = [0,1]$

$\Omega = \mathbb{R}$

$\Omega = \mathbb{R}^k$

$\mathscr{E} = \{\emptyset, [0,0.5], (0.5,1.0], [0,1]\}$

Typically: $\mathscr{E} = B(\Omega)$ ("Borel field")

Borel field is the set of all subsets of non-negligible size (e.g., intervals $[0.1, 0.1 + \epsilon]$)

# Discrete vs. Continuous Sample Spaces

**Discrete (countable) outcomes**

$\Omega = \{1,2,3,4,5,6\}$

$\Omega = \{\text{person, robot, camera, TV}, \ldots\}$

$\Omega = \mathbb{N}$

$\mathscr{E} = \{\emptyset, \{1,2\}, \{3,4,5,6\}, \{1,2,3,4,5,6\}\}$

Typically**:** $\mathscr{E} = \mathscr{P}(\Omega)$

**Question:**
$\mathscr{E} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$?

**Continuous (uncountable) outcomes**

$\Omega = [0,1]$

$\Omega = \mathbb{R}$

$\Omega = \mathbb{R}^k$

$\mathscr{E} = \{\emptyset, [0,0.5], (0.5,1.0], [0,1]\}$

Typically: $\mathscr{E} = B(\Omega)$ ("Borel field")

**Note:** *not* $\mathscr{P}(\Omega)$

# Exercise

- Write down the power set of {1, 2, 3}
- More advanced: Why is the power set a valid event space? Hint: Check the two properties

**Definition:**

A non-empty set $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ is an **event space** if it satisfies

1. $A \in \mathscr{E} \implies A^c \in \mathscr{E}$
2. $A_1, A_2, \ldots \in \mathscr{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathscr{E}$

# Exercise answer

A set $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ is an **event space** if it satisfies

1. $A \in \mathscr{E} \implies A^c \in \mathscr{E}$
2. $A_1, A_2, \ldots \in \mathscr{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathscr{E}$

- $\Omega = \{1,2,3\}$
- $\mathscr{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$
- Proof that the power set satisfies the two properties
- Take any $A \in \mathscr{P}(\Omega)$ (e.g., $A = \{1\}$ or $A = \{1,2\}$). Then $A^c = \Omega \backslash A$ is a subset of $\Omega$, and so $A^c \in \mathscr{P}(\Omega)$ since the power set contains all subsets
- Take any $A, B \in \mathscr{P}(\Omega)$. Then $A \cup B \subset \Omega$, and so $A \cup B \in \mathscr{P}(\Omega)$
- More generally, for an infinite union, see: https://proofwiki.org/wiki/Power_Set_is_Closed_under_Countable_Unions

# Axioms

**Definition:**

Given a measurable space $(\Omega, \mathscr{E})$, any function $P : \mathscr{E} \to [0,1]$ satisfying

1. **unit measure:** $P(\Omega) = 1$, and
2. **$\sigma$-additivity:** $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for any countable sequence $A_1, A_2, \ldots \in \mathscr{E}$ where $A_i \cap A_j = \emptyset$ whenever $i \neq j$

is a **probability measure** (or **probability distribution**).

If $P$ is a probability measure over $(\Omega, \mathscr{E})$, then $(\Omega, \mathscr{E}, P)$ is a **probability space**.

# Defining a Distribution

**Example:**

$\Omega = \{0,1\}$

$\mathscr{E} = \{\emptyset, \{0\}, \{1\}, \Omega\}$

$$P = \begin{cases} 1 - \alpha & \text{if } A = \{0\} \\ \alpha & \text{if } A = \{1\} \\ 0 & \text{if } A = \emptyset \\ 1 & \text{if } A = \Omega \end{cases}$$

where $\alpha \in [0,1]$.

**Questions:**

1. Do you recognize this distribution?
2. How should we choose $P$ in practice?
   a. Can we choose an arbitrary function?
   b. How can we guarantee that all of the constraints will be satisfied?

We will define distributions using **PMFs** and **PDFs**

# Probability Mass Functions (PMFs)

**Definition:** Given a **discrete** sample space $\Omega$ and event space $\mathscr{E} = \mathscr{P}(\Omega)$, any function $p : \Omega \to [0,1]$ satisfying $\sum_{\omega \in \Omega} p(\omega) = 1$ is a **probability mass function**.

- For a discrete sample space, instead of defining $P$ directly, we can define a **probability mass function** $p : \Omega \to [0,1]$.
- $p$ gives a probability for **outcomes** instead of **events**
- The probability for any event $A \in \mathscr{E}$ is then defined as $P(A) = \sum_{\omega \in A} p(\omega)$.

# Example: PMF for a Fair Die

A **categorical distribution** is a distribution over a **finite** outcome space, where the probability of each outcome is specified separately.

**Example: Fair Die**

$$\Omega = \{1,2,3,4,5,6\}$$

$$p(\omega) = \frac{1}{6}$$

$$p(\{3,4\}) = \frac{1}{3}$$

| $\omega$ | $p(\omega)$ |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |

**Questions:**

1. What is a possible event? What is its probability?
2. What is the event space?

# Moving to Boolean Terminology with Random Variables

**Example:** Suppose we observe both a die's number, and where it lands.

$\Omega = \{(left,1), (right,1), (left,2), (right,2), \ldots, (right,6)\}$

We might want to think about the probability that we get a large number, without thinking about where it landed.

Let $X$ = number that comes up. We could ask about $P(X = 3)$ or $P(X \geq 4)$

This is simpler to write than using the event notation, e.g,

$P(X = 3)$ would be written $P(\{\omega \in \Omega \mid \omega_2 = 3\})$

# Random Variables, Formally

Given a probability space $(\Omega, \mathscr{E}, P)$, a **random variable** is a function $X : \Omega \to \mathscr{X}$ (where $\mathscr{X}$ is a new outcome space), satisfying

$$\{\omega \in \Omega \mid X(\omega) \in A\} \in \mathscr{E} \quad \forall A \in B(\mathscr{X}).$$

It follows that $P_X(A) = P(\{\omega \in \Omega \mid X(\omega) \in A\})$.

**Example:** Let $\Omega$ be a population of people, $\omega = (\text{height, age, \ldots, location})$, and $X(\omega)$ = height in cm, and the event $A = [150,170]$.

$$P(X \in A) = P(150 \leq X \leq 170) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

# RVs are intuitive

- All the probability rules remain the same, since RVs are a mapping to create a new outcome space, event space and probabilities
- The notation may look onerous, but they simply formalize something we do naturally: specify the variable we care about, knowing it is defined by a more complex underlying distribution
- We have really already been talking about RVs
  - e.g., for $X =$ dice outcome, event $A = \{5,6\}$, $P(A) = P(X \geq 4)$

# Random Variables Simplify Terminology

- A Boolean expression involving random variables defines an event:

  E.g., $P(X \geq 4) = P(\{\omega \in \Omega \mid X(\omega) \geq 4\})$

- Random variables strictly generalize the way we can talk about probabilities
  - lets us be specific about any transformations
  - switches language from events to boolean expressions
- From this point onwards, we will exclusively reason in terms of random variables

# Revisiting the Fair Die PMF

**Example: Fair Die**

$\mathcal{X} = \{1,2,3,4,5,6\}$

$$p(x) = \frac{1}{6}$$

| $x$ | $p(x)$ |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |

**Questions:**

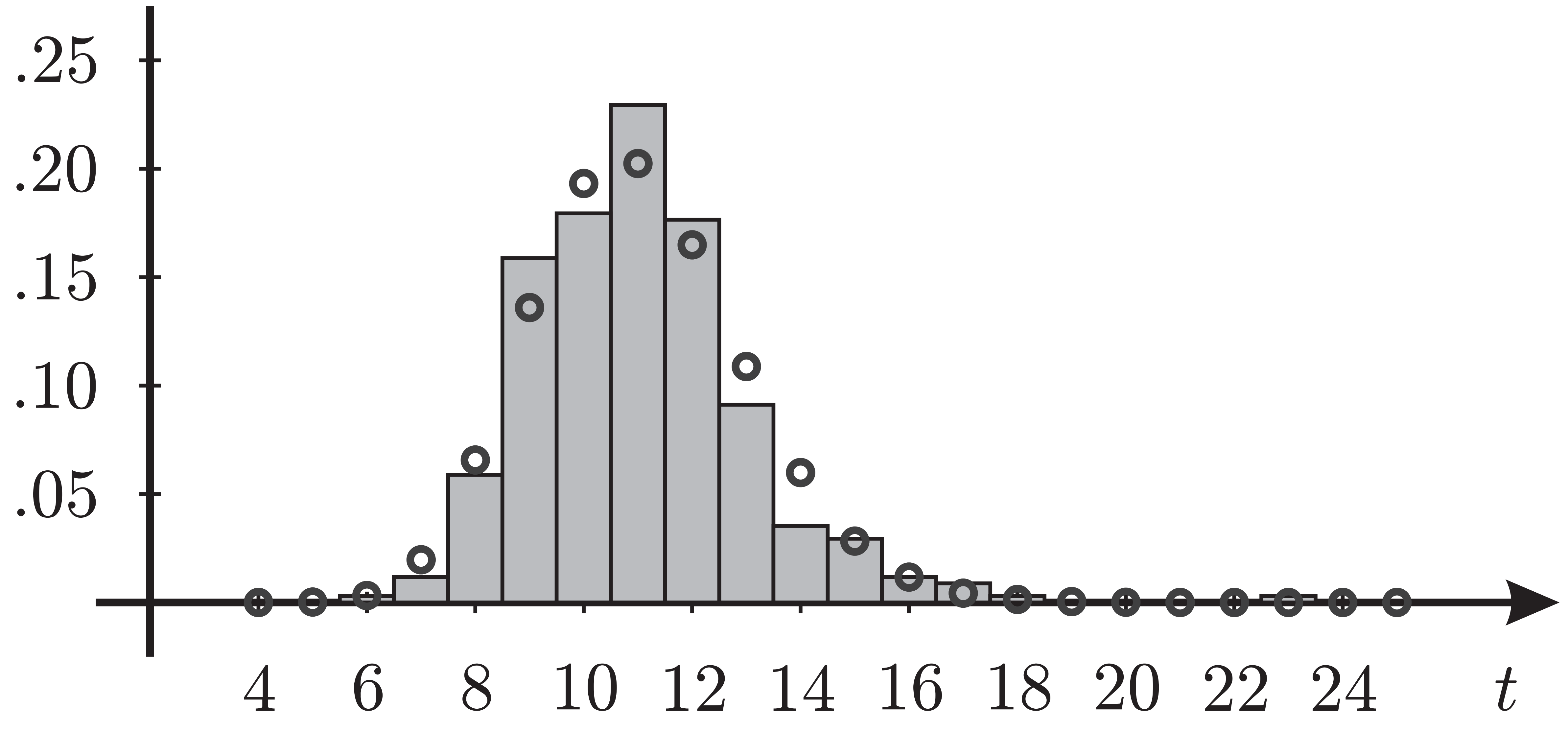
1. What is a possible event? What is its probability?
2. What is the event space?

Answer: event space and probabilities are the same, but we write the probabilistic question using booleans

$$p(\{3,4\}) = \frac{1}{3} \longrightarrow p(3 \leq X \leq 4) = \frac{1}{3}$$ or $$p(X \in \{3,4\}) = \frac{1}{3}$$

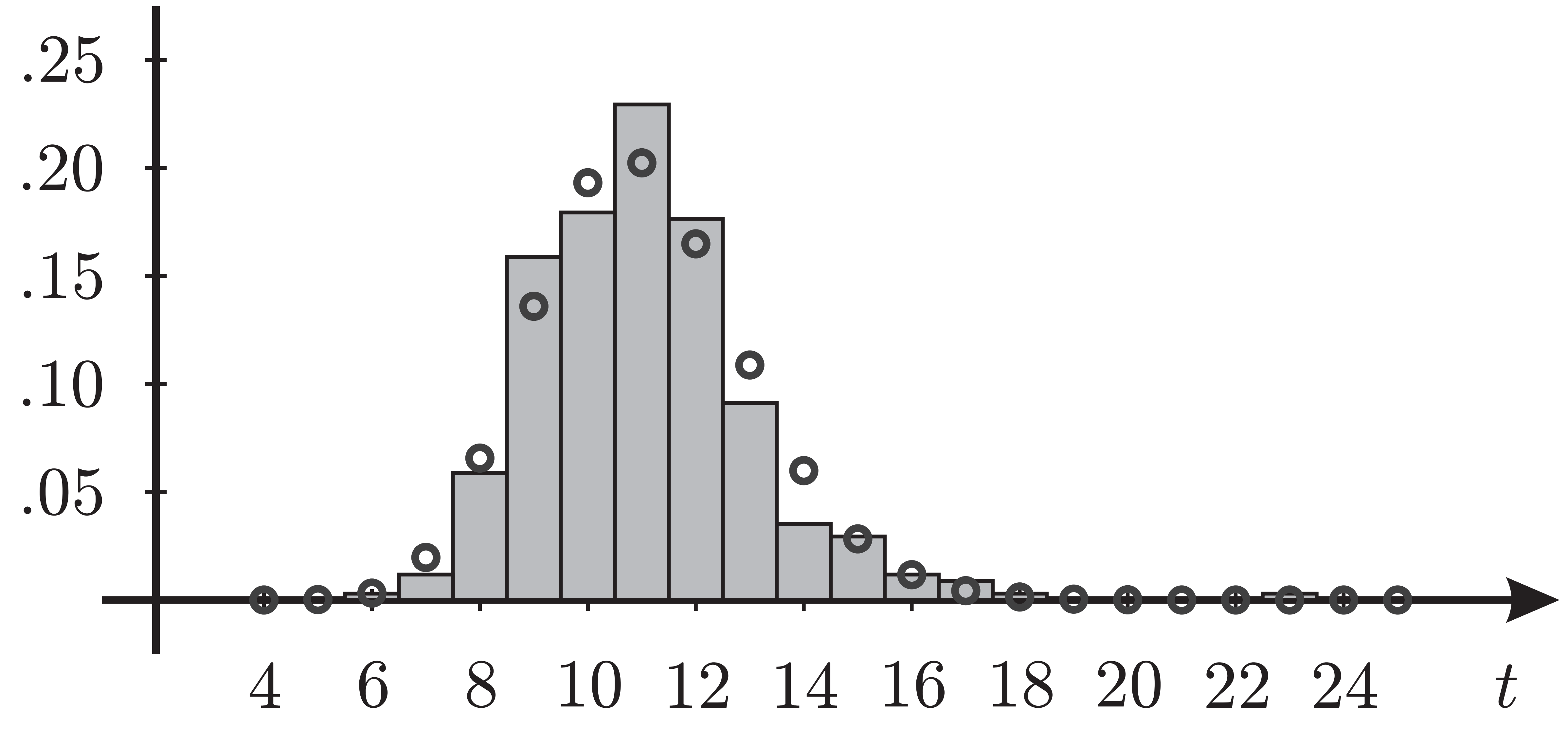
# Example: Using a PMF

- Suppose that you recorded your commute time (in minutes) every day for a year (i.e., 365 recorded times).
- The random variable is $T$ with outcomes $t \in \{4,5,6,7,\ldots,25\}$
- **Question:** How do you get $p(t)$?
- **Question:** How is $p(t)$ useful?
- **Question:** How do you compute $p(10 \leq T \leq 13)$?



# Example: Using a PMF

- Suppose that you recorded your commute time (in minutes) every day for a year (i.e., 365 recorded times).
- The random variable is $T$ with outcomes $t \in \{4,5,6,7,\ldots,25\}$
- **Question:** How do you get $p(t)$? (Answer: count and normalize)
- **Question:** How is $p(t)$ useful?
  - We can take mode as prediction
- **Question:** How do you compute $p(10 \leq T \leq 13)$?
  - Answer: $\sum_{t \in \{10,11,12\}} p(t)$




This PMF is called a categorical distribution, with 21 categories (table of probabilities)

# Useful PMFs: Bernoulli

A **Bernoulli distribution** is a special case of a **categorical distribution** in which there are only two outcomes. It has a single **parameter** $\alpha \in (0,1)$.

$\mathcal{X} = \{T, F\}$ (or $\mathcal{X} = \{S, F\}$)

$$p(x) = \begin{cases} \alpha & \text{if } x = T \\ 1 - \alpha & \text{if } x = F \end{cases}.$$

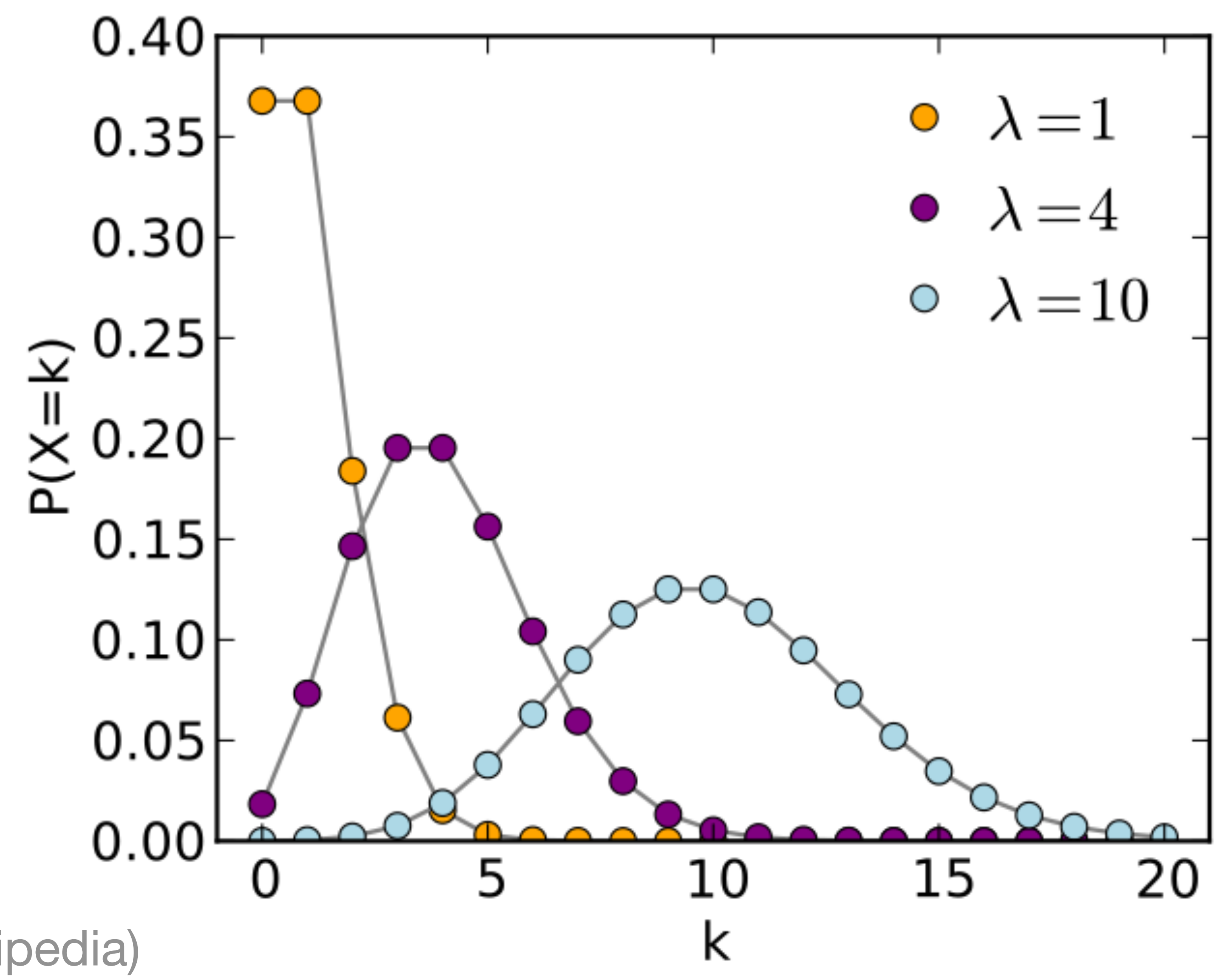
Alternatively: $\mathcal{X} = \{0,1\}$

$p(x) = \alpha^x (1-\alpha)^{1-x}$ for $x \in \{0,1\}$

# Useful PMFs: Poisson

A **Poisson distribution** is a distribution over the non-negative integers.
It has a single parameter $\lambda \in (0, \infty)$.

E.g., number of calls received by a call centre in an hour, $\lambda$ is the average number of calls



(Image: Wikipedia)

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

**Questions:**

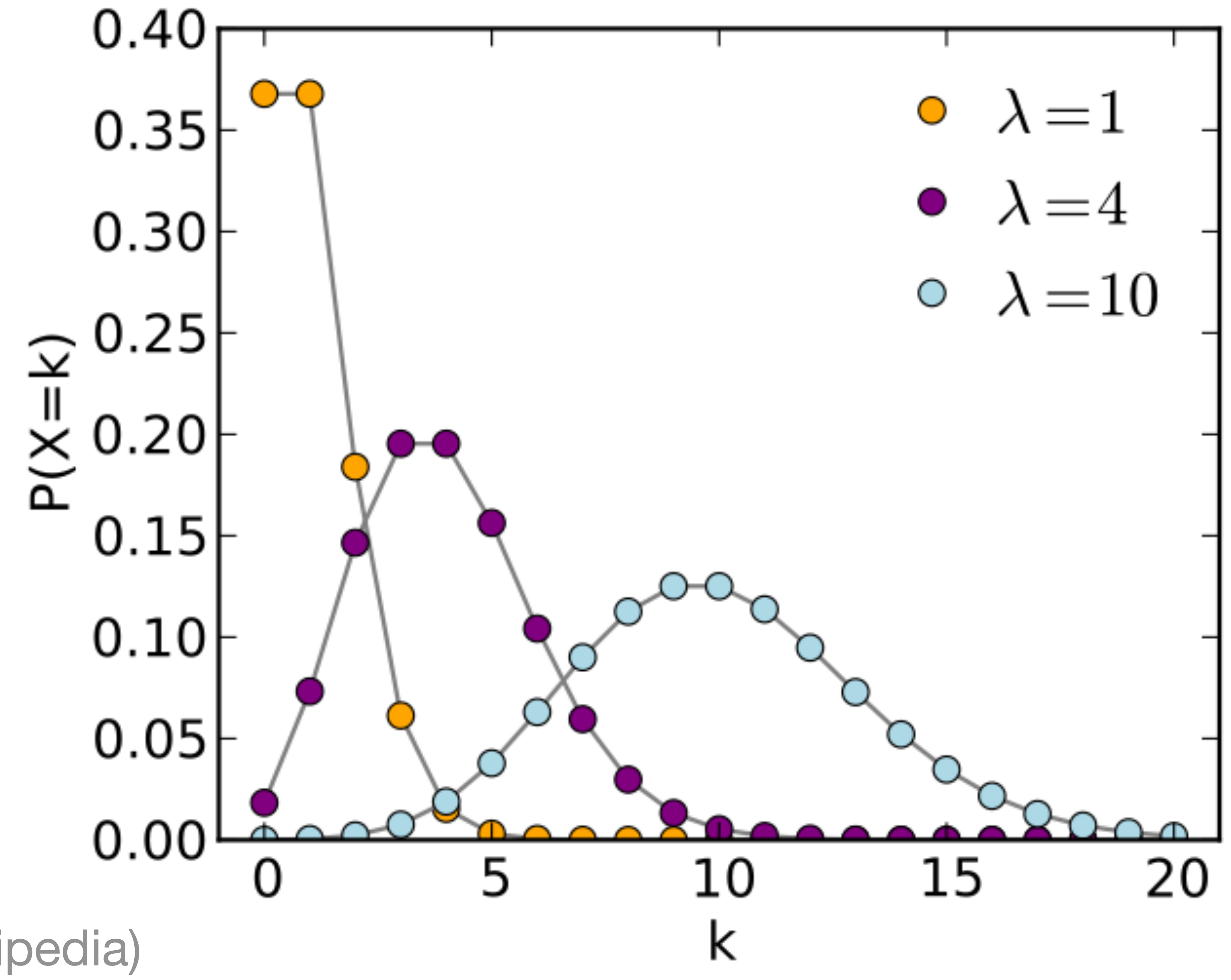
1. Could we define this with a table instead of an equation?
2. How can we check whether this is a valid PMF?
3. $\lambda$ real-valued, but outcomes are discrete. What might be the mode (most likely outcome)?

# Useful PMFs: Poisson

A **Poisson distribution** is a distribution over the non-negative integers.

It has a single parameter $\lambda \in (0, \infty)$.

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$




(Image: Wikipedia)

1. Could we define this with a table instead of an equation?
   - No because the outcome space is infinite

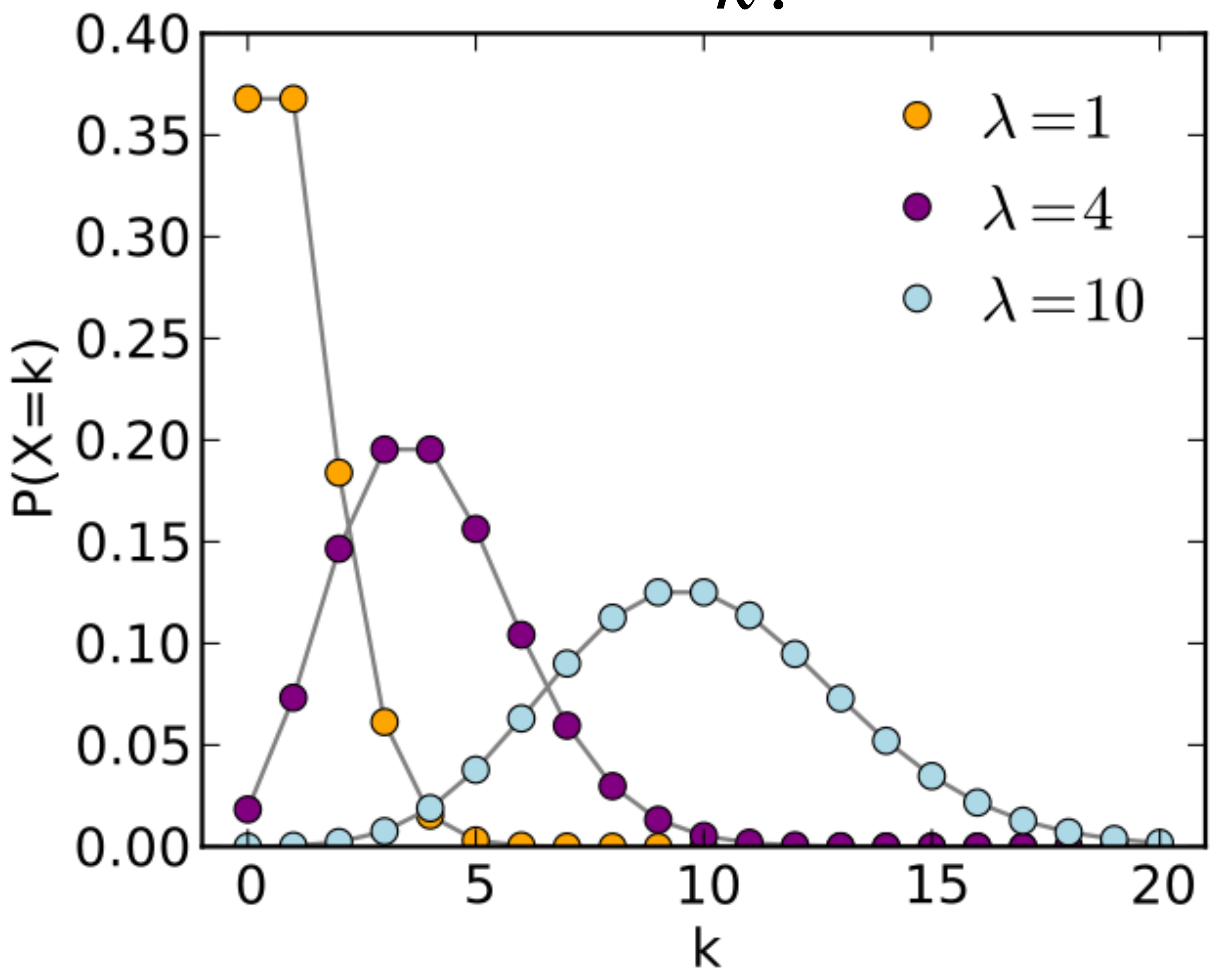
2. How can we check whether this is a valid PMF?
   - Check if $\sum_{k=0}^{\infty} p(k) = 1$

3. $\lambda$ real-valued, but outcomes are discrete. What might be the mode (most likely outcome)?
   - Mean is $\lambda$, may not correspond to any outcome
   - Two modes, $\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$

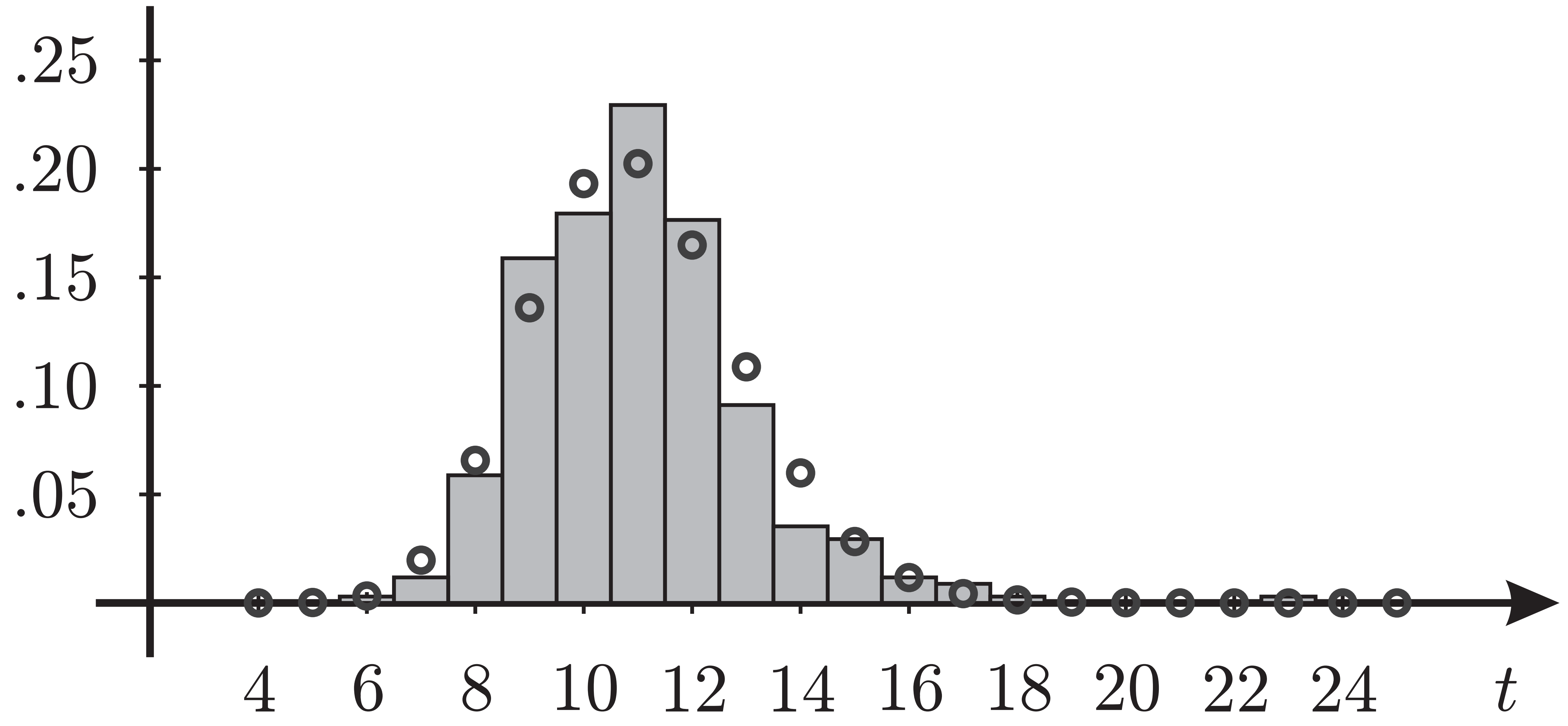
# Commute Times Again

- **Question:** Could we use a **Poisson distribution** for commute times (instead of a categorical distribution)?
- **Question:** What would be the benefit of using a Poisson distribution? Hint: what do you need to estimate to specify the Poisson, vs the categorical?

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

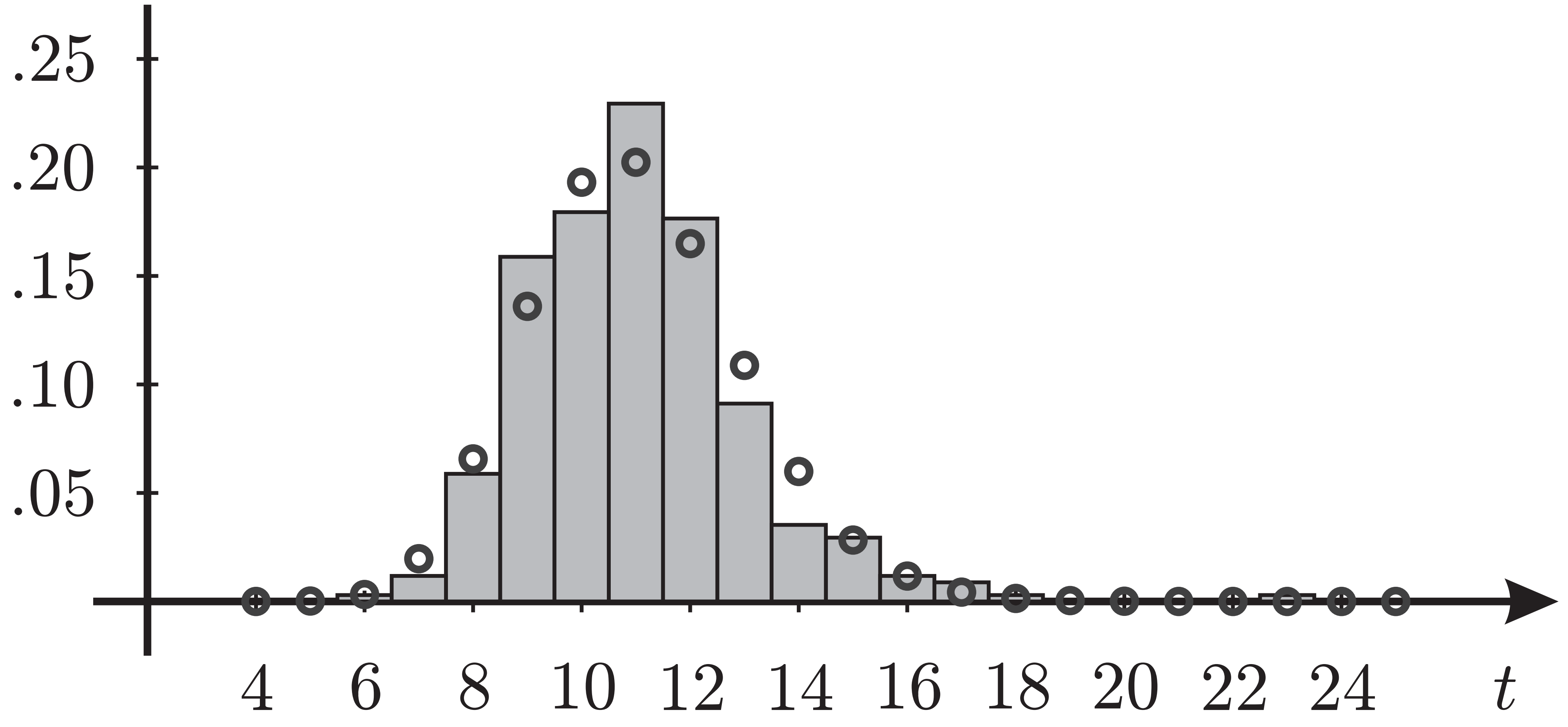



$$p(4) = 1/365,\ p(5) = 2/365,\ p(6) = 4/365,\ \ldots$$



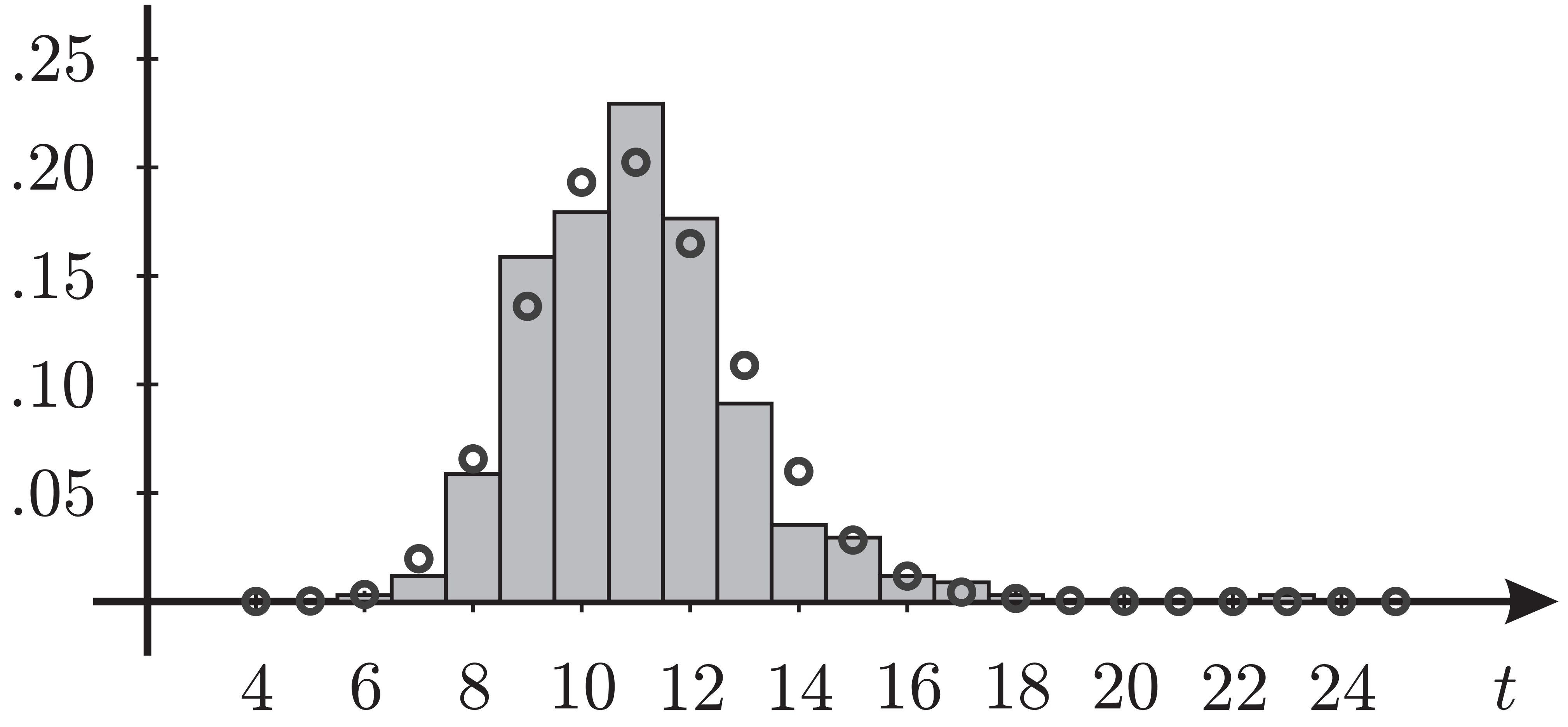
# Continuous Commute Times

- It never actually takes *exactly* 12 minutes; I rounded each observation to the nearest integer number of minutes.
  - Actual data was 12.345 minutes, 11.78213 minutes, etc.



# Continuous Commute Times

- It never actually takes *exactly* 12 minutes; I rounded each observation to the nearest integer number of minutes.
  - Actual data was 12.345 minutes, 11.78213 minutes, etc.
- **Question:** Could we use a Poisson distribution to predict the *exact* commute time (rather than the nearest number of minutes)? Why?



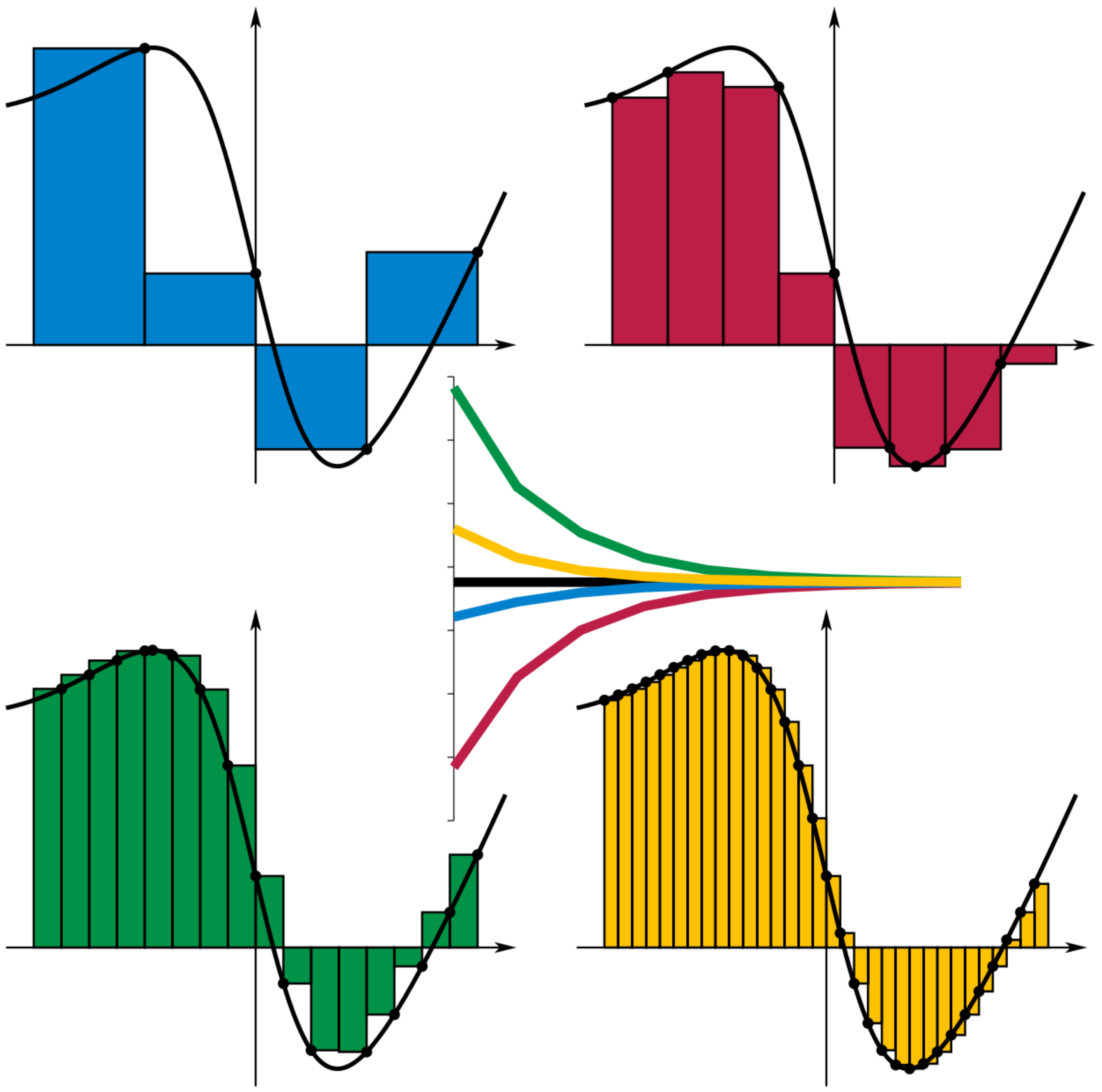
# Probability Density Functions (PDFs)

**Definition:** Given a continuous sample space $\mathscr{X}$ and event space $\mathscr{E} = B(\mathscr{X})$, any function $p : \mathscr{X} \to [0,\infty)$ satisfying $\displaystyle\int_{\mathscr{X}} p(x)dx = 1$ is a **probability density function**.

- For a continuous sample space, instead of defining $P$ directly, we can define a **probability density function** $p : \mathscr{X} \to [0,\infty)$.
- The probability for any event $A \in \mathscr{E}$ is then defined as

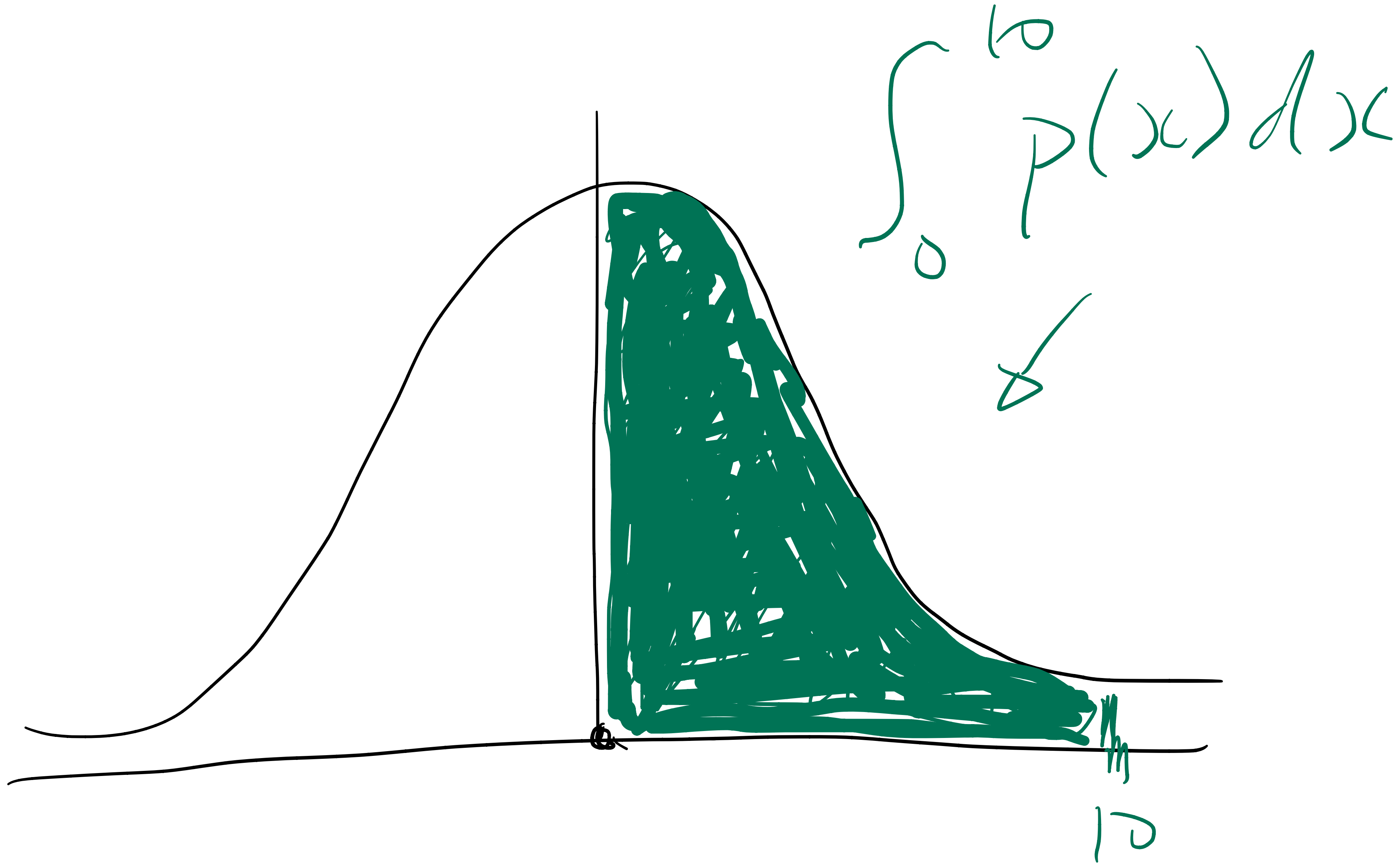
$$P(A) = \int_A p(x)dx.$$

# Recall Integration



# Integration to give the probability of an event

- Imagine the PDF looks like the following concave function



$$p(0 \leq X \leq 10) = \int_0^{10} p(x)dx$$

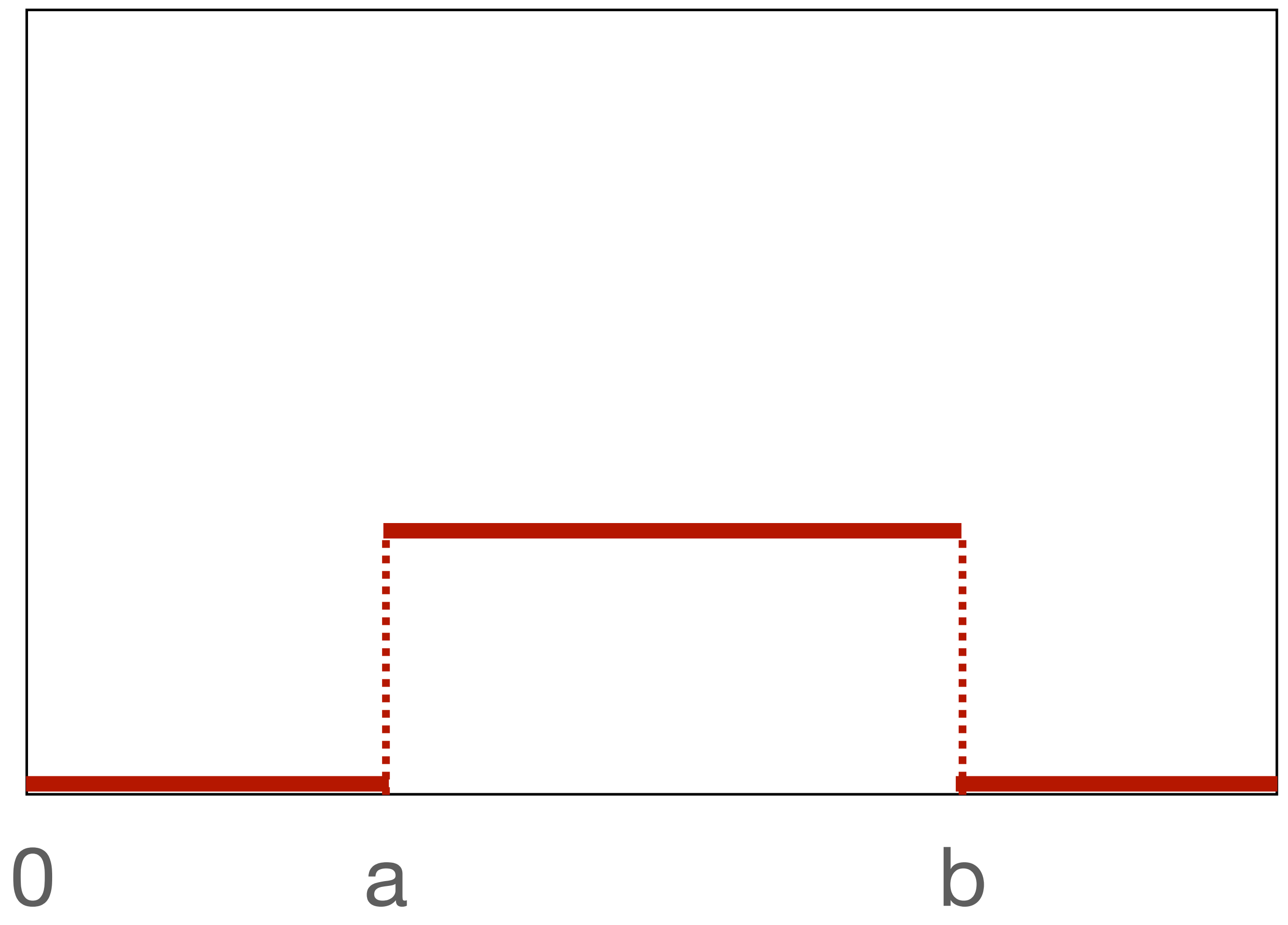
Area under the curve reflects the probability of seeing an outcome in that region

# Useful PDFs: Uniform

A **uniform distribution** is a distribution over a real interval. It has two parameters: $a$ and $b$.

$$\mathcal{X} = [a, b]$$

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$




**Question**: Does $\mathcal{X}$ have to be bounded?

# Exercise: Check that the uniform pdf satisfies the required properties

Recall that the antiderivative of 1 is x, because the derivative of x is 1

$$\int_a^b p(x)dx = \int_a^b \frac{1}{b-a}dx$$

$$= \frac{1}{b-a}\int_a^b dx = \frac{1}{b-a}x\,|_a^b$$

$$= \frac{1}{b-a}(b-a) = 1$$

# Useful PDFs: Gaussian

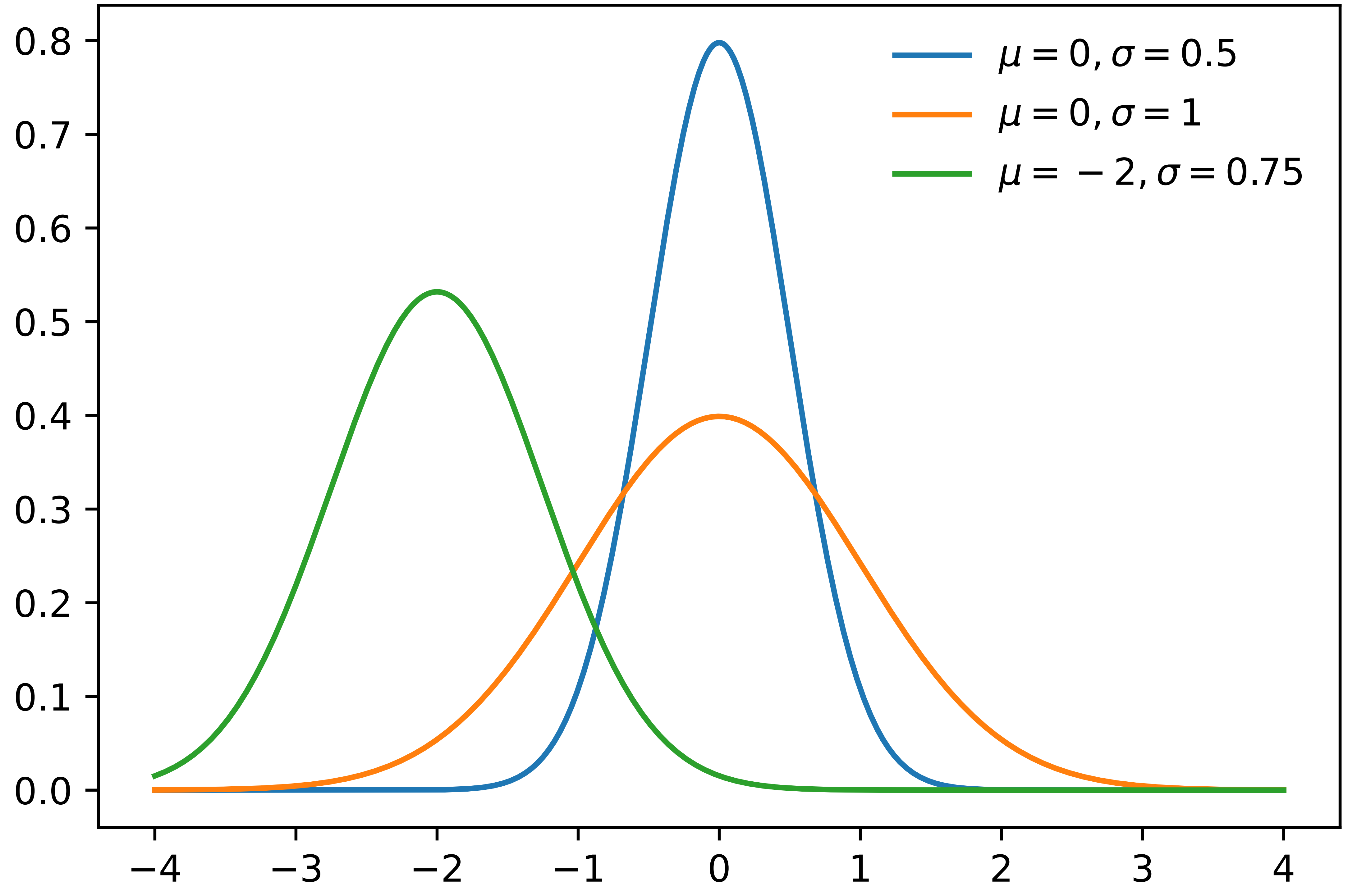
A **Gaussian distribution** is a distribution over the real numbers. It has two parameters: $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

$$\mathcal{X} = \mathbb{R}$$

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

where $\exp(x) = e^x$

Also called a normal distribution and written $\mathcal{N}(\mu, \sigma^2)$



# Why the distinction between PMFs and PDFs?

1. When the sample space $\mathcal{X}$ is **discrete:**
   - Singleton event: $P(\{x\}) = p(x)$ for $x \in \mathcal{X}$

$$P(A) = \sum_{x \in \mathcal{X}} p(x)$$

2. When the sample space $\mathcal{X}$ is **continuous:**
   - Example: Stopping time for a car with $\mathcal{X} = [3,12]$
   - **Question:** What is the probability that the stopping time is *exactly* 3.14159?

$$P(A) = \int_A p(x)dx$$

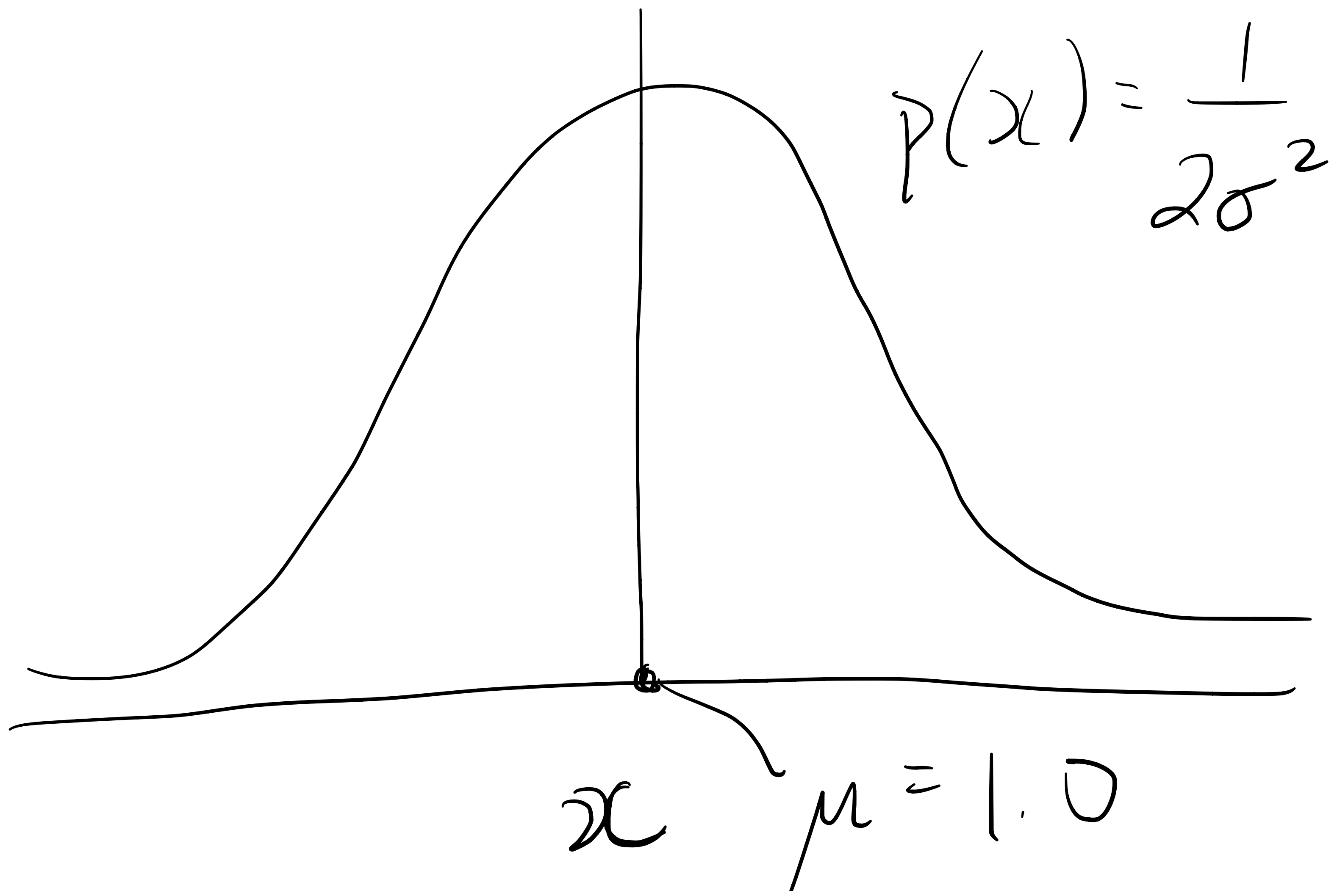
$$P(\{3.14159\}) = \int_{3.14159}^{3.14159} p(x)dx = 0$$

   - More reasonable: Probability that stopping time is between 3 to 3.5.

# Example comparing integration and summation

Imagine we have a Gaussian distribution

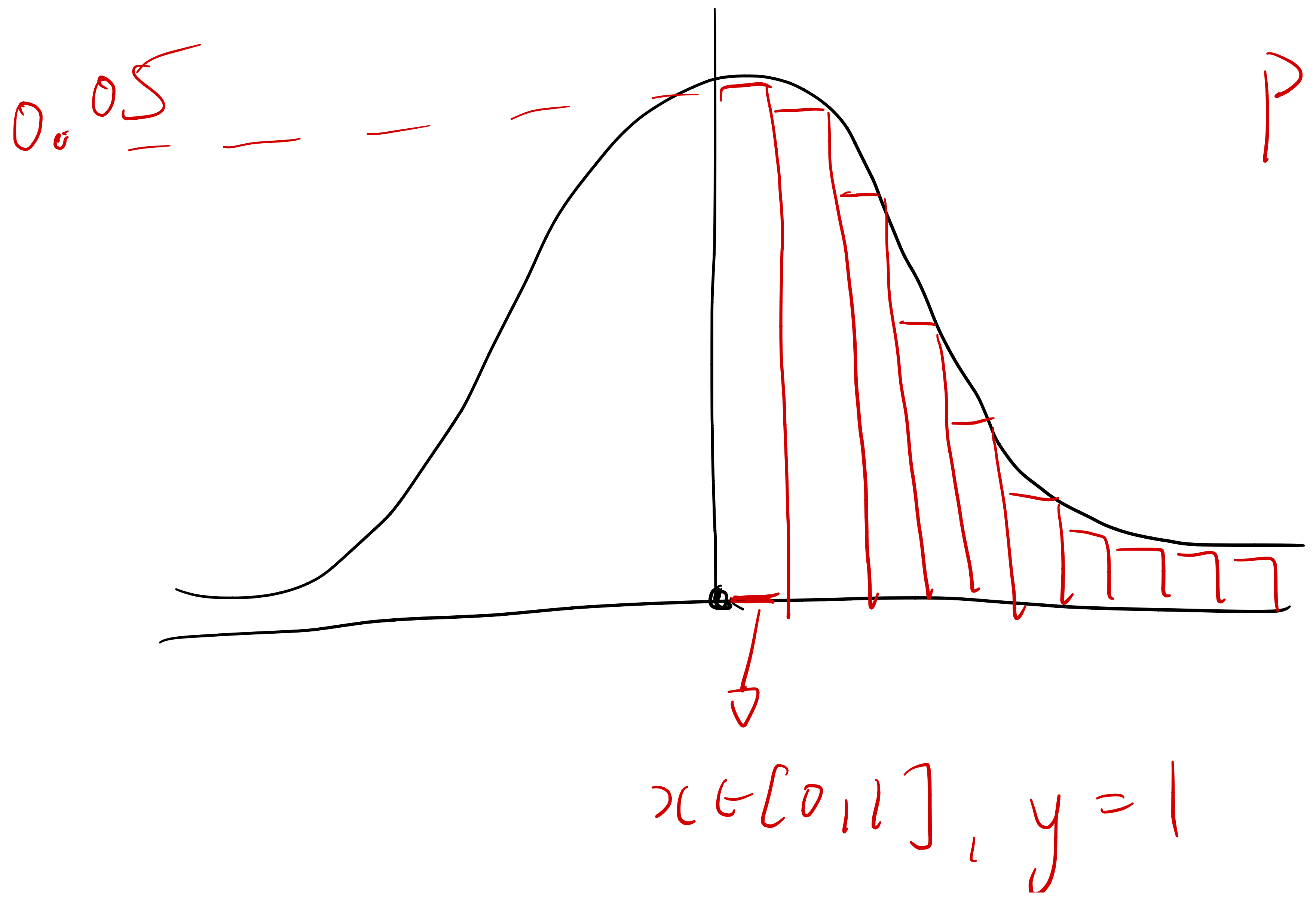
$$p(x) = \frac{1}{2\sigma^2} \exp\left( - \frac{(x-\mu)^2}{2\sigma^2} \right)$$



# Example comparing integration and summation (cont)

Let's pretend we discretized to get a PMF

$y = i$ for $x \in (i-1, i]$

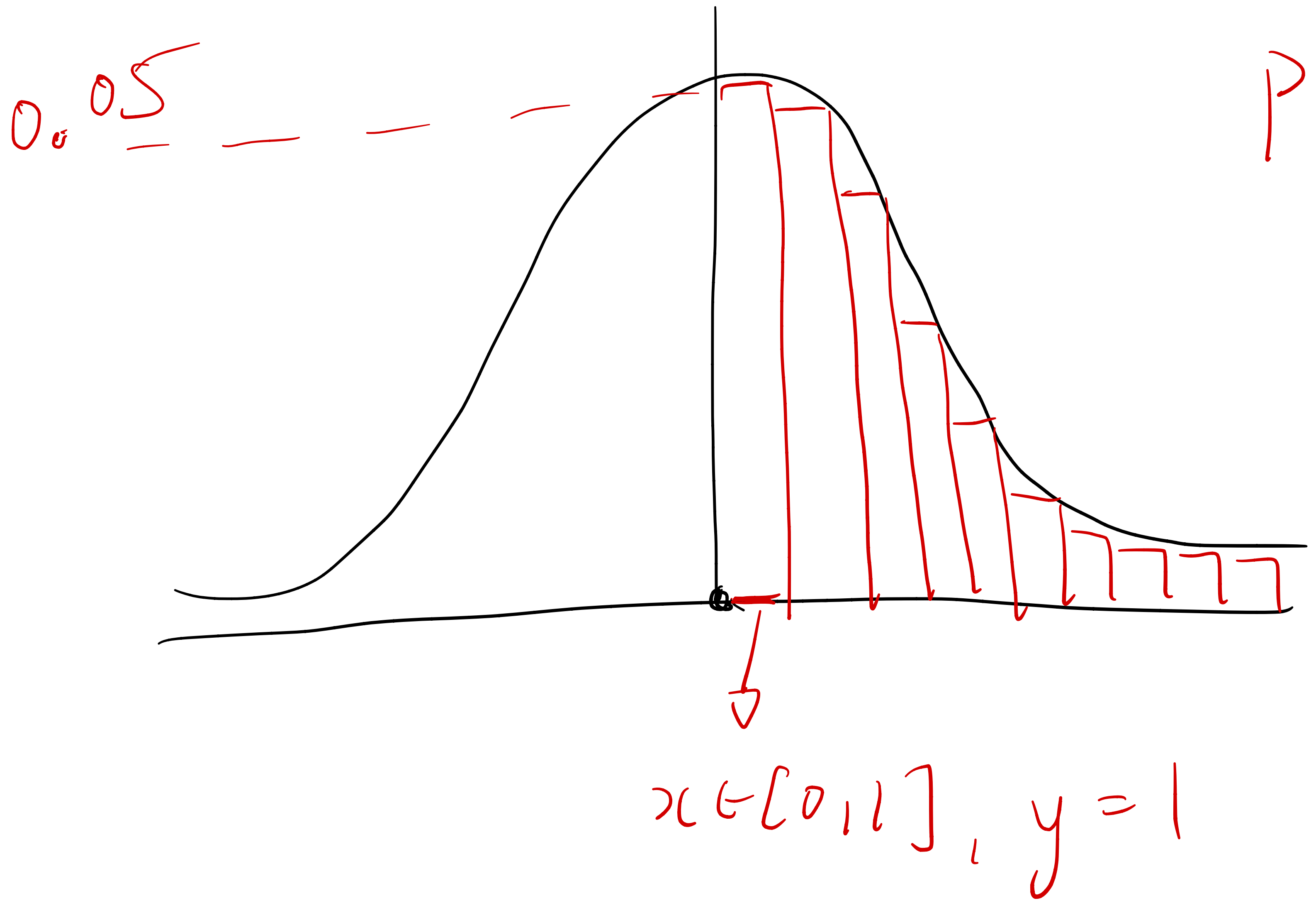



$p(y=1) = 0.05$

# Example comparing integration and summation (cont)

Let's pretend we discretized to get a PMF

$y = i$ for $x \in (i-1, i]$




$p(y=1) = 0.05$

when we ask

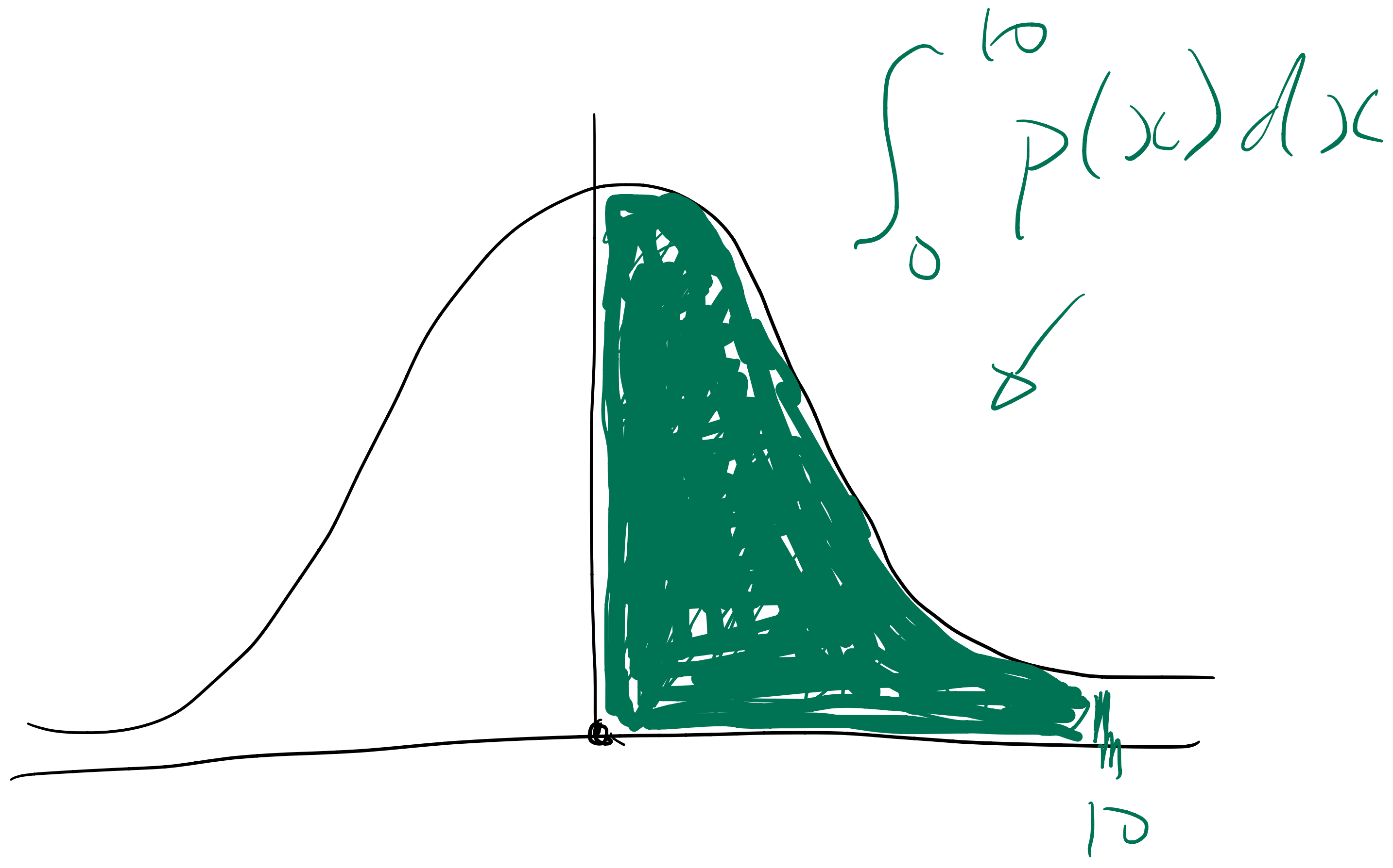
$$\Pr(X \in [0,10]) = \int_0^{10} p(x)\,dx$$

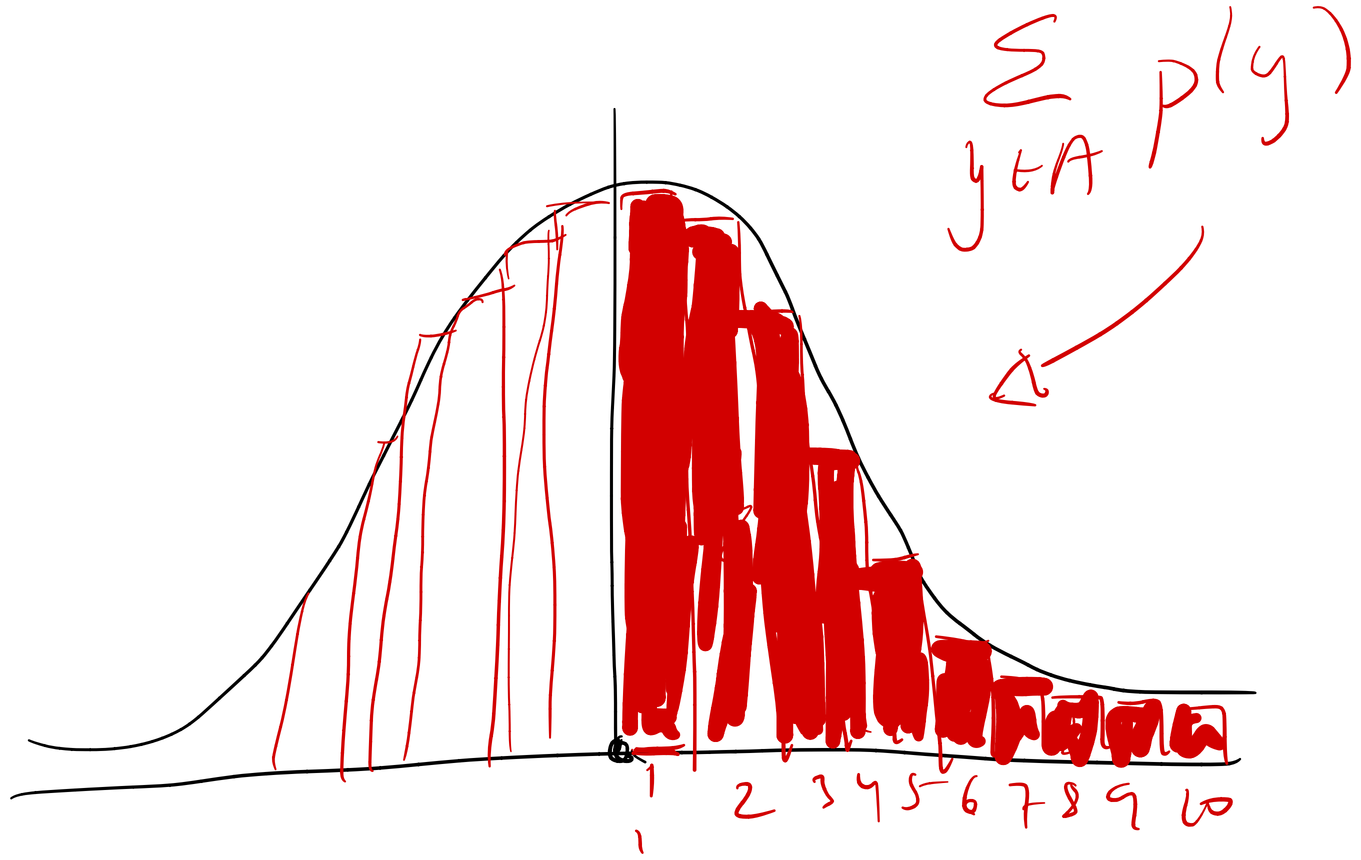
Similar to

$$\Pr(Y \in \underbrace{\{1, 2, 3, \ldots, 10\}}_{A}) = \sum_{y \in A} p(y)$$

# Example comparing integration and summation (cont)

Both reflect density or mass in a region.




Note: technically the red rectangles should go a bit above the Gaussian line, if we really did discretize.
My drawing is not perfect here.

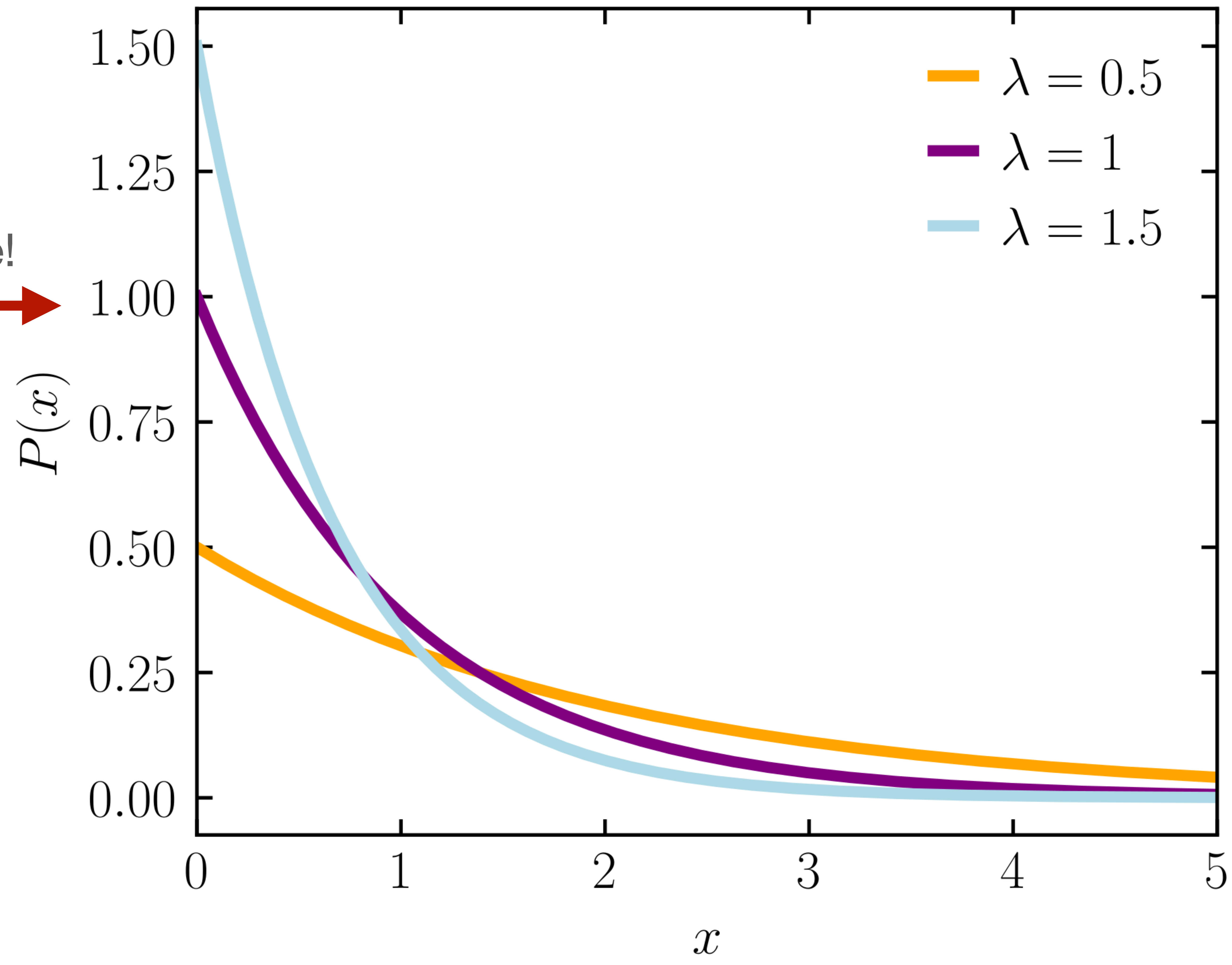
# Useful PDFs: Exponential

An **exponential distribution** is a distribution over the positive reals. It has one parameter $\lambda > 0$.

$$\mathscr{X} = \mathbb{R}^+$$

$$p(x) = \lambda \exp(-\lambda x)$$

1 is here!



# Why can the density be above 1?

Consider an interval event $A = [x, x + \Delta x]$, for small $\Delta x$.

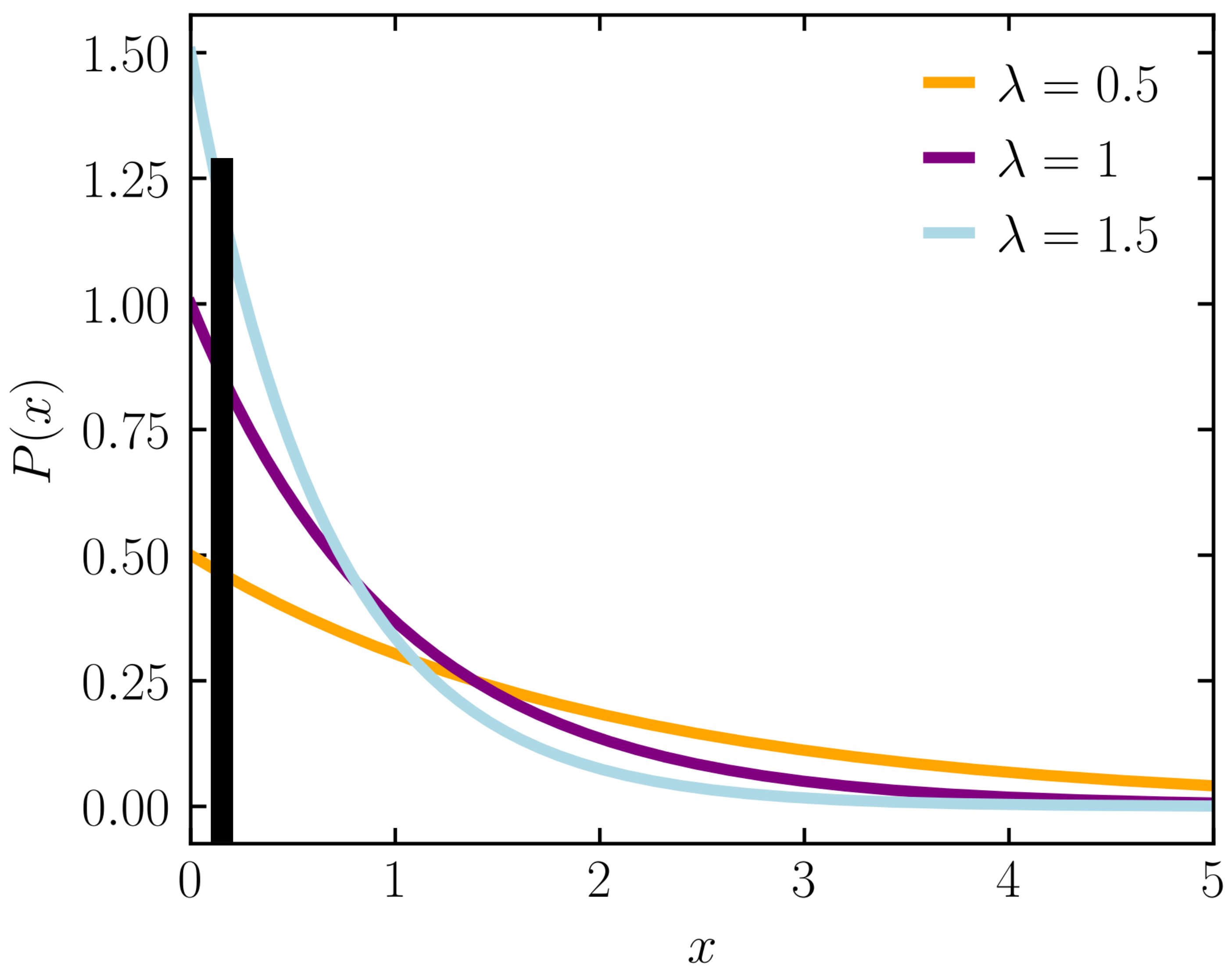
$$P(A) = \int_x^{x+\Delta x} p(x)\,dx$$

e.g., $x = 0.1, \quad \Delta x = 0.01 \quad p(x) = 1.5\exp(-1.5x),\ p(0.1) \approx 1.3$

$$\approx p(x)\Delta x$$

$$p(X \in [0.1, 0.11]) \approx 1.3 \times 0.01 = 0.013$$

- $p(x)$ can be big, because $\Delta x$ can be very small
  - In particular, $p(x)$ can be bigger than 1
- But $P(A)$ **must** be less than or equal to 1



# Exercise

- Imagine I asked you to tell me the probability that my birthday is on February 10 or July 9.
  - What is the outcome space and what is the event for this question?
  - Would we use a PMF or PDF to model these probabilities?
- Imagine I asked you to tell me the probability that the Uber would be here in between 3-5 minutes
  - What is the outcome space and what is the event for this question?
  - Would we use a PMF or PDF to model these probabilities?

# Summary

- Probabilities are a means of **quantifying uncertainty**
- A probability distribution is defined on a measurable space consisting of a **sample space** and an **event space**.
- **Discrete** sample spaces (and random variables) are defined in terms of **probability mass functions** (PMFs)
- **Continuous** sample spaces (and random variables) are defined in terms of **probability density functions** (PDFs)
- **Random variables** let us reason about probabilistic questions at a more abstract level with boolean expressions