Outline

1. Intro
   - Bayesian vs Frequentist Interpretations

2. Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3. Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Outline

1 Intro
   • Bayesian vs Frequentist Interpretations

2 Probability Theory Review
   • Foundations of Probability
   • Random Variables
   • Discrete Random Variables
   • Important Rules of Probability
   • Independence and Conditional Independence
   • Continuous Random Variables

3 Common Discrete Distributions - Univariate
   • Binomial and Bernoulli Distributions
   • Multinomial and Multinoulli Distributions
   • Poisson Distribution
   • Empirical Distribution
what is probability?
there are actually at least two different interpretations of probability

1 frequentist: probabilities represent long run frequencies of events (trials)
2 Bayesian: probability is used to quantify our uncertainty about something (information rather than repeated trials)

coin toss event:
1 frequentist: if we flip the coin many times, we expect it to land heads about half the time
2 Bayesian: we believe the coin is equally likely to land heads or tails on the next toss

advantage of the Bayesian interpretation: it can be used to model our uncertainty about events that do not have long term frequencies; frequentist needs repetition

the basic rules of probability theory are the same, no matter which interpretation is adopted
In order to define a **probability space** we need 3 components \( \{\Omega, \mathcal{F}, P\} \):

- **sample space** \( \Omega \): the set of all the outcomes of a random **experiment**. Here, each **outcome** (realization) \( \omega \in \Omega \) can be thought of as a *complete description of the state of the real world* at the end of the experiment.

- **event space** \( \mathcal{F} \): a set whose elements \( A \in \mathcal{F} \) (called **events**) are subsets of \( \Omega \). (i.e., \( A \subseteq \Omega \) is a collection of possible outcomes of an experiment) \( \mathcal{F} \) should satisfy 3 properties (\( \sigma \)-algebra of events):
  1. \( \emptyset \in \mathcal{F} \)
  2. \( A \in \mathcal{F} \Rightarrow \overline{A} = \Omega \setminus A \in \mathcal{F} \) (closure under complementation)
  3. \( A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F} \) (closure under countable union)

- **probability measure** \( P \): a function \( P : \mathcal{F} \rightarrow \mathbb{R} \) that satisfies the following 3 axioms of probability:
  1. \( P(A) \geq 0 \) for all \( A \in \mathcal{F} \)
  2. \( P(\Omega) = 1 \)
  3. if \( A_1, A_2, \ldots \) are disjoint events (i.e., \( A_i \cap A_j = \emptyset \) whenever \( i \neq j \)), then \( P(\bigcup_i A_i) = \sum_i P(A_i) \) (\( P \) is countably additive)
A simple example

experiment: tossing a six-sided dice
- sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ (a simple representation)
- trivial event space
  - $\mathcal{F} = \{\emptyset, \Omega\}$
  - unique probability measure satisfying the requirements is given by
    $P(\emptyset) = 0, \ P(\Omega) = 1$
- power set event space
  - $\mathcal{F} = 2^{\Omega}$ (i.e., the set of all subsets of $\Omega$)
  - a possible probability measure
    $P(i) = 1/6$ for $i \in \{1, 2, 3, 4, 5, 6\} = \Omega$

question: do the above sample space outcomes completely describe the state of a dice-tossing experiment?
some **important properties** on events (can be inferred from axioms)

- \( A \subseteq B \Rightarrow P(A) \leq P(B) \)
- \( P(A \cap B) \leq \text{min}(P(A), P(B)) \)
- union bound: \( P(A \cup B) \leq P(A) + P(B) \)
- complement rule: \( P(\overline{A}) = P(\Omega \setminus A) = 1 - P(A) \)
- impossible event: \( P(\emptyset) = 0 \)
- law of total probability: if \( A_1, ..., A_k \) are a set of disjoint events such that \( \bigcup_{i=1}^{N} A_i = \Omega \) then \( \sum_{i=1}^{N} P(A_i) = 1 \)

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

1 events can be represented by using Venn diagrams
Conditional Probability

- let $B$ be an event with non-zero probability, i.e. $p(B) > 0$
- the **conditional probability** of any event $A$ given $B$ is defined as
  \[
P(A|B) = \frac{p(A \cap B)}{p(B)}\]
- in other words, $P(A|B)$ is the probability measure of the event $A$ after observing the occurrence of event $B$
- two events are called **independent** iff
  \[
P(A \cap B) = P(A)P(B) \quad (\text{or equivalently } P(A|B) = P(A))\]
- therefore, **independence** is equivalent to saying that observing $B$ does not have any effect on the probability of $A$
Conditional Probability

**a frequentist intuition of conditional probability**

- $N$ is total number of experiment trials
- For an event $E$, let’s define $P(E) \triangleq \frac{N_E}{N}$ where $N_E$ is the number of trials where $E$ is verified

Hence for events $A$ and $B$ (considering the limit $N \to \infty$)

- $P(A) = \frac{N_A}{N}$ where $N_A$ is the number of trials where $A$ is verified
- $P(B) = \frac{N_B}{N}$ where $N_B$ is the number of trials where $B$ is verified
- $P(A \cap B) = \frac{N_{A\wedge B}}{N}$ where $N_{A\wedge B}$ is the number of trials where both $A$ and $B$ are verified

Let’s consider only the trials where $B$ is verified, hence

- $P(A|B) = \frac{N_{A\wedge B}}{N_B}$ \hspace{1cm} ($N_B > 0$ now acts as $N$)
- Dividing by $N$, one obtains $P(A|B) = \frac{N_{A\wedge B}/N}{N_B/N} = \frac{P(A\cap B)}{P(B)}$
Outline

1 Intro
- Bayesian vs Frequentist Interpretations

2 Probability Theory Review
- Foundations of Probability
- Random Variables
  - Discrete Random Variables
  - Important Rules of Probability
  - Independence and Conditional Independence
  - Continuous Random Variables

3 Common Discrete Distributions - Univariate
- Binomial and Bernoulli Distributions
- Multinomial and Multinoulli Distributions
- Poisson Distribution
- Empirical Distribution
Random Variables

intuition: a random variable represents an interesting "aspect" of the outcomes $\omega \in \Omega$

more formally:

- a random variable $X$ is a function $X : \Omega \to \mathbb{R}$
- a random variable is denoted by using upper case letters $X(\omega)$ or more simply $X$ (here $X$ is a function)
- the particular values (instances) of a random variable may take on are denoted by using lower case letters $x$ (here $x \in \mathbb{R}$)

types of random variables:

- discrete random variable: function $X(\omega)$ can only take values in a finite set $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ or countably infinite set (e.g. $\mathcal{X} = \mathbb{N}$)
- continuous random variable: function $X(\omega)$ can take continuous values in $\mathbb{R}$
Random Variables

A random variable is a measurable function

- Since \( X(\omega) \) takes values in \( \mathbb{R} \), let's try to define an "event space" on \( \mathbb{R} \): in general we would like to observe if \( X(\omega) \in B \) for some subset \( B \subset \mathbb{R} \).

- As "event space" on \( \mathbb{R} \), we can consider \( B \) the Borel \( \sigma \)-algebra on the real line\(^2\), which is generated by the set of half-lines \( \{ (-\infty, a] : a \in (-\infty, \infty) \} \) by repeatedly applying union, intersection and complement operations.

- An element \( B \subset \mathbb{R} \) of the Borel \( \sigma \)-algebra \( \mathcal{B} \) is called a Borel set.

- The set of all open/closed subintervals in \( \mathbb{R} \) are contained in \( \mathcal{B} \).

- For instance, \((a, b) \in \mathcal{B}\) and \([a, b] \in \mathcal{B}\).

- A random variable is a measurable function \( X : \Omega \to \mathbb{R} \), i.e.

\[
X^{-1}(B) = \{ \omega \in \Omega : X(\omega) \in B \} \in \mathcal{F} \quad \text{for each } B \in \mathcal{B}
\]

i.e., if we consider an "event" \( B \in \mathcal{B} \) this can be represented by a proper event \( F_B \in \mathcal{F} \) where we can apply the probability measure \( P \).

\(^2\)Here we should use the notation \( \mathcal{B}(\mathbb{R}) \), for simplicity we drop \( \mathbb{R} \).
we have defined the probability measure $P$ on $\mathcal{F}$, i.e. $P : \mathcal{F} \to \mathbb{R}$

how to define the probability measure $P_X$ w.r.t. $X$?

$$P_X(B) \triangleq P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

which is well-defined given that $X^{-1}(B) \in \mathcal{F}$

at this point, we have an induced probability space

$$\{\Omega_X, \mathcal{F}_X, P_X\} \triangleq \{\mathbb{R}, \mathcal{B}, P_X\}$$

and we can equivalently reason on it
1 Intro
   - Bayesian vs Frequentist Interpretations

2 Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3 Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Discrete Random Variables

**discrete Random Variable (RV)**

- $X(\omega)$ can only take values in a finite set $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ or in a countably infinite set
- how to define the probability measure $P_X$ w.r.t. $X$?

$$P_X(X = x_k) \triangleq P(\{\omega : X(\omega) = x_k\})$$

- in this case $P_X$ returns measure one to a countable set of reals
- a simpler way to represent the probability measure is to directly specify the probability of each value the discrete RV can assume
- in particular, a **Probability Mass Function** (PMF) is a function $p_X : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$p_X(X = x) \triangleq P_X(X = x)$$

- it's very common to drop the subscript $X$ and denote the PMF with $p(X) = p_X(X = x)$
Outline

1. Intro
   - Bayesian vs Frequentist Interpretations

2. Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - **Important Rules of Probability**
     - Independence and Conditional Independence
     - Continuous Random Variables

3. Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Important Rules of Probability

considering two discrete RV $X$ and $Y$ at the same time

- **sum rule**
  
  \[ p(X) = \sum_Y p(X, Y) \]  
  (marginalization)

- **product rule**
  
  \[ p(X, Y) = p(X|Y)p(Y) \]

- **chain rule:**
  
  \[ p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)...p(X_D|X_{1:D-1}) \]

  where $1:D$ denotes the set $\{1, 2, ..., D\}$  
  (Matlab-like notation)
Important Rules of Probability

a frequentist intuition of the sum rule

- $N$ number of trials
- $n_{ij}$ number of trials in which $X = x_i$ and $Y = y_j$
- $c_i$ number of trials in which $X = x_i$, one has $c_i = \sum_j n_{ij}$
- $r_j$ number of trials in which $Y = y_j$, one has $r_j = \sum_i n_{ij}$
- $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$ (considering the limit $N \to \infty$)

hence:

- $p(X = x_i) = \frac{c_i}{N} = \sum_j \frac{n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j)$
Bayes’ Theorem

combining the definition of condition probability with the product and sum rules:

1. \( p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \)  
   (conditional prob. def.)

2. \( p(X, Y) = p(Y \mid X)p(X) \) 
   (product rule)

3. \( p(Y) = \sum_X p(X, Y) = \sum_X p(Y \mid X)p(X) \) 
   (sum rule + product rule)

one obtains the **Bayes’ Theorem**

\[
p(X \mid Y) = \frac{p(Y \mid X)p(X)}{\sum_X p(Y \mid X)p(X)}
\]

N.B.: we could write \( p(X \mid Y) \propto p(Y \mid X)p(X) \); the denominator \( p(Y) = \sum_X p(Y \mid X)p(X) \) 

can be considered as a normalization constant
Bayes’ Theorem
An Example

events:
- $C$ = breast cancer present, $\overline{C}$ = no cancer
- $M$ = positive mammogram test, $\overline{M}$ = negative mammogram test

probabilities:
- $p(C) = 0.4\%$ (hence $p(\overline{C}) = 1 - p(C) = 99.6\%$)
- if there is cancer, the probability of a pos mammogram is $p(M|C) = 80\%$
- if there is no cancer, we still have $p(M|\overline{C}) = 10\%$

the false conclusion: positive mammogram $\Rightarrow$ the person is 80% likely to have cancer

question: what is the conditional probability $p(C|M)$?

$$p(C|M) = \frac{p(M|C)p(C)}{p(M)} = \frac{p(M|C)p(C)}{p(M|C)p(C) + p(M|\overline{C})p(\overline{C})}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

the true conclusion: positive mammogram $\Rightarrow$ the person is about 3% likely to have cancer
Outline

1 Intro
   · Bayesian vs Frequentist Interpretations

2 Probability Theory Review
   · Foundations of Probability
   · Random Variables
   · Discrete Random Variables
   · Important Rules of Probability
   · Independence and Conditional Independence
   · Continuous Random Variables

3 Common Discrete Distributions - Univariate
   · Binomial and Bernoulli Distributions
   · Multinomial and Multinoulli Distributions
   · Poisson Distribution
   · Empirical Distribution
Independence and Conditional Independence

considering two RV $X$ and $Y$ at the same time

- $X$ and $Y$ are **unconditionally independent**

  $$X \perp Y \iff p(X, Y) = p(X)p(Y)$$

  in this case $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$

- $X_1, X_2, \ldots, X_D$ are **mutually independent** if

  $$p(X_1, X_2, \ldots, X_D) = p(X_1)p(X_2)\ldots p(X_D)$$

- $X$ and $Y$ are **conditionally independent**

  $$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z)$$

  in this case $p(X|Y, Z) = p(X|Z)$
Outline

1 Intro
   • Bayesian vs Frequentist Interpretations

2 Probability Theory Review
   • Foundations of Probability
   • Random Variables
   • Discrete Random Variables
   • Important Rules of Probability
   • Independence and Conditional Independence
   • Continuous Random Variables

3 Common Discrete Distributions - Univariate
   • Binomial and Bernoulli Distributions
   • Multinomial and Multinoulli Distributions
   • Poisson Distribution
   • Empirical Distribution
Continuous Random Variables

A continuous random variable

- $X(\omega)$ can take any value on $\mathbb{R}$
- how to define the probability measure $P_X$ w.r.t. $X$?

$$P_X(X \in B) \triangleq P(X^{-1}(B)) \quad (\text{with } B \in \mathcal{B})$$

- in this case $P_X$ gives zero measure to every singleton set, and hence to every countable set\(^3\)

---

\(^3\) unless we consider some particular/degenerate cases
given a continuous RV $X$

- **Cumulative Distribution Function** (CDF): $F(x) \triangleq P_X(X \leq x)$
  
  - $0 \leq F(x) \leq 1$
  - the CDF is a monotonically non-decreasing
    $F(x) \leq F(x + \Delta x)$ with $\Delta x > 0$
  - $F(-\infty) = 0$, $F(\infty) = 1$
  - $P_X(a < X \leq b) = F(b) - F(a)$

- **Probability Density Function** (PDF): $p(x) \triangleq \frac{dF}{dx}$
  
  - we assume $F$ is continuous and the derivative exists
  - $F(x) = P_X(X \leq x) = \int_{-\infty}^{x} p(\xi) d\xi$
  - $P_X(x < X \leq x + dx) \approx p(x) dx$
  - $P_X(a < X \leq b) = \int_{a}^{b} p(x) dx$

$p(x)$ acts as a density in the above computations
Some Properties

reconsider

1. \( P_X(a < X \leq b) = \int_a^b p(x) \, dx \)

2. \( P_X(x < X \leq x + dx) \approx p(x) \, dx \)

- the first implies \( \int_{-\infty}^{\infty} p(x) \, dx = 1 \) (consider \( (a, b) = (-\infty, \infty) \))
- the second implies \( p(x) \geq 0 \) for all \( x \in \mathbb{R} \)
- it is possible that \( p(x) > 1 \), for instance, consider the uniform distribution with PDF

\[
\text{Unif}(x|a, b) = \frac{1}{b - a} \mathbb{1}(a \leq x \leq b)
\]

if \( a = 0 \) and \( b = 1/2 \) then \( p(x) = 2 \) in \( [a, b] \)
- Assume $F$ is continuous (this was required for defining $p(x)$).
- We have that $P_X(X = x) = 0$ (zero probability on a singleton set).
- In fact for $\epsilon \geq 0$:
  
  $$P_X(X = x) \leq P_X(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon) = \delta F(x, \epsilon)$$

  and given that $F$ is continuous $P_X(X = x) \leq \lim_{\epsilon \to 0} \delta F(x, \epsilon) = 0$. 

given that the CDF $F$ is monotonically increasing, let’s consider its inverse $F^{-1}$

$$F^{-1}(\alpha) = x_\alpha \iff P_X(X \leq x_\alpha) = \alpha$$

$x_\alpha$ is called the $\alpha$ quantile of $F$

$F^{-1}(0.5)$ is the median

$F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the lower and upper quartiles

for symmetric PDFs (e.g. $\mathcal{N}(0, 1)$) we have $F^{-1}(1 - \alpha/2) = -F^{-1}(\alpha/2)$ and the central interval $(F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2))$ contains $1 - \alpha$ of the mass probability
Mean and Variance

• **mean or expected value** $\mu$

  for a discrete RV:  
  $$\mu = \mathbb{E}[X] \triangleq \sum_{x \in \chi} x \cdot p(x)$$

  for a continuous RV:  
  $$\mu = \mathbb{E}[X] \triangleq \int_{x \in \chi} x \cdot p(x) \, dx$$  
  (defined if $\int_{x \in \chi} |x| \cdot p(x) \, dx < \infty$)

• **variance** $\sigma^2 = \text{var}[X] \triangleq \mathbb{E}[(X - \mu)^2]$

  $$\text{var}[X] = \mathbb{E}[(X - \mu)^2] = \int_{x \in \chi} (x - \mu)^2 \cdot p(x) \, dx =$$

  $$= \int_{x \in \chi} x^2 \cdot p(x) \, dx - 2\mu \int_{x \in \chi} x \cdot p(x) \, dx + \mu^2 \int_{x \in \chi} p(x) \, dx = \mathbb{E}[X^2] - \mu^2$$

  (this can be also obtained for discrete RV)

• **standard deviation** $\sigma = \text{std}[X] = \sqrt{\text{var}[X]}$
Moments

- $n$-th moment

for a discrete RV: $\mathbb{E}[X^n] \triangleq \sum_{x \in \chi} x^n p(x)$

for a continuos RV: $\mathbb{E}[X^n] \triangleq \int_{x \in \chi} x^n p(x) \, dx$ (defined if $\int_{x \in \chi} |x|^n p(x) \, dx < \infty$)
Outline

1. Intro
   - Bayesian vs Frequentist Interpretations

2. Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3. Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Binomial Distribution

- we toss a **coin** $n$ times
- $X$ is a discrete RV with $x \in \{0, 1, \ldots, n\}$, the occurred number of heads
- $\theta$ is the probability of heads
- $X \sim \text{Bin}(n, \theta)$ i.e., $X$ has a **binomial distribution** with PMF

$$\text{Bin}(k|n, \theta) \triangleq \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad (= P_X(X = k))$$

where we use the **binomial coefficient**

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

- mean $= n\theta$
- var $= n\theta(1 - \theta)$

N.B.: recall that $(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$
we toss a **coin** only one time

$X$ is a discrete RV with $x \in \{0, 1\}$ where $1 =$ head, $0 =$ tail

$\theta$ is the probability of heads

$X \sim \text{Ber}(\theta)$ i.e., $X$ has a **Bernoulli distribution** with PMF

$$\text{Ber}(x|\theta) \triangleq \theta^{I(x=1)}(1-\theta)^{I(x=0)}$$

that is

$$\text{Ber}(x|\theta) = \begin{cases} 
\theta & \text{if } x = 1 \\
1-\theta & \text{if } x = 0 
\end{cases}$$

- mean $= \theta$
- var $= \theta(1-\theta)$
Outline

1. Intro
   - Bayesian vs Frequentist Interpretations

2. Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3. Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
we toss a $K$-sided dice $n$ times

the possible outcome is $x = (x_1, x_2, ..., x_K)$ where $x_j \in \{0, 1, ..., n\}$ is the number of times side $j$ occurred

$n = \sum_{j=1}^{K} x_j$

$\theta_j$ is the probability of having side $j$

$\sum_{j=1}^{K} \theta_j = 1$

$X \sim \text{Mu}(n, \theta)$ i.e., $X$ has a multinomial distribution with PMF

$$\text{Mu}(x | n, \theta) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

where we use the multinomial coefficient

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \ldots x_K!}$$

which is the num of ways to divide a set of size $n$ into subsets of size $x_1, x_2, ..., x_K$
we toss the **dice** only one time

the possible outcome is \( x = (\mathbb{I}(x_1 = 1), \mathbb{I}(x_2 = 1), \ldots, \mathbb{I}(x_K = 1)) \) where \( x_j \in \{0, 1\} \) represents if side \( j \) occurred or not (dummy encoding or one-hot encoding)

\( \theta_j \) is the probability of having side \( j \), i.e., \( p(x_j = 1|\theta) = \theta_j \)

\( X \sim \text{Cat}(\theta) \) i.e., \( X \) has the **categorical distribution** (or **multinoulli**)

\[
\text{Cat}(x|\theta) = \text{Mu}(x|1, \theta) \triangleq \prod_{j=1}^{K} \theta_j^{x_j}
\]
1 Intro
   - Bayesian vs Frequentist Interpretations

2 Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3 Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Poisson Distribution

- $X$ is a discrete RV with $x \in \{0, 1, 2, \ldots\}$ (support on $\mathbb{N}^+$)
- $X \sim \text{Poi}(\lambda)$ i.e., $X$ has a **Poisson distribution** with PMF
  \[
Poi(x|\lambda) \equiv e^{-\lambda} \frac{\lambda^x}{x!}
\]
  recall that $e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$

- this distribution is used as a model for counts of rare events (e.g. accidents, failures, etc)
1. Intro
   - Bayesian vs Frequentist Interpretations

2. Probability Theory Review
   - Foundations of Probability
   - Random Variables
   - Discrete Random Variables
   - Important Rules of Probability
   - Independence and Conditional Independence
   - Continuous Random Variables

3. Common Discrete Distributions - Univariate
   - Binomial and Bernoulli Distributions
   - Multinomial and Multinoulli Distributions
   - Poisson Distribution
   - Empirical Distribution
Empirical Distribution

given a dataset \( D = \{x_1, x_2, \ldots, x_N\} \)

the \textbf{empirical distribution} is defined as

\[
p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)
\]

\( 0 \leq w_i \leq 1 \) are the weights

\[
\sum_{i=1}^{N} w_i = 1
\]

\( \delta_{x_i}(x) = \mathbb{I}(x = x_i) \)

this can be view as an \textbf{histogram} with "spikes" at \( x_i \in D \) and 0-probability out \( D \)
Credits

- Kevin Murphy’s book
- G. Chandalia "A gentle introduction to Measure Theory"