Probability, Random Variables, Distributions

Keng-Chi Chang

Department of Political Science
University of California San Diego

2023-09-19

Building Blocks

- Outcome: A specific result
  - One coin flip: Outcome is either H or T
  - Two coins flipped in sequence: HT is an outcome
- **Sample space**: The *set* of all possible outcomes, often denoted S
  - One coin flip: $S_1 = \{H, T\}$
  - Two coins flipped in sequence: $S_2 = \{HH, HT, TH, TT\}$
- **Events**: *Subset* of possible outcomes, a subset of S (can be S itself).
  - $A_1 = \{H\} \subseteq S_1$; $A_2 = \{HH, HT\} \subseteq S_2$
- **Probability**: Chance of an event within the sample space
  - A function that maps events to $[0, 1]$
  - Will give a more formal definition

Different Sizes of Sets

- Roughly, intuition tells us that

$$\text{Probability} = \frac{\text{Size of Event}}{\text{Size of Sample Space}} = \frac{\text{\# Outcomes}}{\text{\# All possible outcomes}}$$

- Sample space can be of different sizes, leading to different treatments of probability
- **Finite**:
  - Ex. Number of seats in US House, $S = \{0, 1, \dots, 435\}$
- Infinite but **Countable**:
  - Ex. Potential number of wars, $S = \{0, 1, 2, 3, \dots\}$
- Infinite and **Uncountable**:
  - Ex. Time duration of cabinets, $S = [0, \infty)$
  - You cannot "count" the number of outcomes in this case

Set Operations

Given two sets A and B, we can do the following operations (draw Venn diagrams):

1. **Union**: The set containing all of the elements in A **or** B

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$$

2. **Intersection**: The set containing all of the elements in both A **and** B

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$$

3. **Complement**: The set containing all of the elements **not** in A

$$A^c = \{\omega : \omega \notin A\}$$

Other useful concepts:

- **Set Difference**: $B \setminus A = \{\omega : \omega \in B \text{ and } \omega \notin A\} = B \cap A^c$
- **Disjoint Sets**: A and B are disjoint if $A \cap B = \emptyset$
- **Indicator Function**: $\mathbb{1}(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$

Probability Function

Definition

A **function** $\mathbb{P}$ which assigns events to $\mathbb{R}$ is called a **probability function** if it satisfies the following Axioms of Probability (Kolmogorov 1933)

1. For any event A, $\mathbb{P}[A] \geq 0$ (non-negative)
2. $\mathbb{P}[S] = 1$ (sum up to 1, normalization)
3. If $A_1, A_2, \ldots$ are mutually disjoint, then (disjoint $\Rightarrow$ sum)

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$$

- Axiom 3 imposes that probabilities are additive on disjoint events

$$\text{If } A \cap B = \emptyset, \text{ then } \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

- Ex: $S = \{H, T\}$, $\mathbb{P}[H] = 0.6$, $\mathbb{P}[T] = 0.6$ is not a valid probability function

Probability Properties

For events A and B, given the three axioms, we can show the following properties:

1. $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$
2. $\mathbb{P}[\varnothing] = 0$
3. $\mathbb{P}[A] \leq 1$
4. Monotone Probability Inequality: If $A \subseteq B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$
5. Inclusion-Exclusion Principle:
$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$
6. Boole's Inequality:
$$\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$$
7. Bonferroni's Inequality:
$$\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$$

Joint Probability and Marginalization

- Joint Event $(A \cap B)$ or simply (A, B): The event that both A and B occur
- **Joint Probability**: Probability that joint event (A, B) occurs, denoted

$$\mathbb{P}[A, B]$$

- Example: Flip a coin twice, $\mathbb{P}[H] = 0.6$, A: 1st coin, B: 2nd coin

- **Marginalization**: We can recover the (marginal) probability $\mathbb{P}[X]$ from joint probability $\mathbb{P}[X, Y]$ by summing over every possible values of Y:

$$\mathbb{P}[X] = \sum_y \mathbb{P}[X, Y = y]$$

Conditional Probability

- **Conditional Probability**: The conditional probability $\mathbb{P}[A \mid B]$ is the probability of A **given that** B has occurred

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A, B]}{\mathbb{P}[B]}$$

  - Allows for the inclusion of **other information** B into the calculation of probability of A
  - Can think of B as the new sample space, and re-normalize all probabilities by $\mathbb{P}[B]$
  - Conditional probability is still a valid probability function (satisfies three axioms)
  - This implies the **Product Rule** of probability: joint = conditional * marginal

$$\mathbb{P}[A, B] = \mathbb{P}[A \mid B] \ \mathbb{P}[B]$$

  or more generally (no particular order is needed),

$$\begin{aligned} \mathbb{P}[A, B, C, D, \ldots] &= \mathbb{P}[A] \ \mathbb{P}[B \mid A] \ \mathbb{P}[C \mid A, B] \ \mathbb{P}[D \mid A, B, C] \cdots \\ &= \mathbb{P}[D] \ \mathbb{P}[C \mid D] \ \mathbb{P}[B \mid C, D] \ \mathbb{P}[A \mid B, C, D] \cdots \end{aligned}$$

Independence

- **Independence**: The occurrence or nonoccurrence of either events A and B have no effect on the occurrence or nonoccurrence of the other; they are **unrelated**.
- The following are equivalent definitions for independence:
  1. $\mathbb{P}[A \mid B] = \mathbb{P}[A]$
  2. $\mathbb{P}[B \mid A] = \mathbb{P}[B]$
  3. $\mathbb{P}[A, B] = \mathbb{P}[A]\,\mathbb{P}[B]$
- Conditioning on the event B does not modify the evaluation of probability of A
- Independent events provide **no information** to each other
- So conditional probability = unconditional probability

Conditional Independence

- **Conditional Independence**: If A and B are independent once you know the occurrence of a **third event** C, then we say that A and B are conditionally independent **given** C.
- The following are equivalent definitions for conditional independence:
  1. $\mathbb{P}[A \mid B, C] = \mathbb{P}[A \mid C]$
  2. $\mathbb{P}[B \mid A, C] = \mathbb{P}[B \mid C]$
  3. $\mathbb{P}[A, B \mid C] = \mathbb{P}[A \mid C]\,\mathbb{P}[B \mid C]$
- This is simply the definitions for independence but adding "$[\cdot \mid C]$"
- This is a somewhat weaker condition than independence since we only need the above equality to hold on some subset of sample space involving event C
  - But independence does not imply conditional independence, or vice versa
- This is one of the foundations of causal inference: $\mathbb{P}[Y \mid D, X] = \mathbb{P}[Y \mid X]$
  - Given covariate X, treatment assignment D is independent of potential outcome Y

Probability: Example

- A box contains two coins: a regular coin and one fake two-headed coin ($\mathbb{P}[H] = 1$).
- I choose a coin at random $\mathbb{P}[C] = p$ and toss it twice. Define the following events:
  - A = First coin toss results in an H
  - B = Second coin toss results in an H
  - C = Regular coin has been selected
- Find the following quantities:
  - $\mathbb{P}[A \mid C]$
  - $\mathbb{P}[B \mid C]$
  - $\mathbb{P}[A, B \mid C]$
  - $\mathbb{P}[A]$
  - $\mathbb{P}[B]$
  - $\mathbb{P}[A, B]$
- A and B are conditional independent given C, but A and B are not independent

Bayes Rule

- From the **product rule** we know that

$$\mathbb{P}[A, B] = \mathbb{P}[B \mid A] \;\; \mathbb{P}[A] = \mathbb{P}[A \mid B] \;\; \mathbb{P}[B]$$

- From **marginalization** we also know that

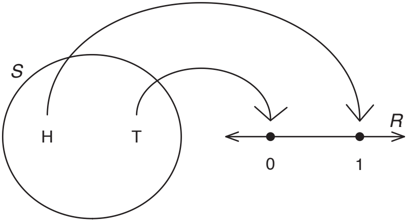
$$\begin{aligned} \mathbb{P}[B] &= \mathbb{P}[A, B] + \mathbb{P}[A^c, B] \\ &= \mathbb{P}[B \mid A] \, \mathbb{P}[A] + \mathbb{P}[B \mid A^c] \, \mathbb{P}[A^c] \end{aligned}$$

- So we can express $\mathbb{P}[A \mid B]$ as a function of $\mathbb{P}[B \mid A]$ (and vise versa):

$$\begin{aligned} \mathbb{P}[A \mid B] = \frac{\mathbb{P}[A, B]}{\mathbb{P}[B]} &= \frac{\mathbb{P}[B \mid A] \, \mathbb{P}[A]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[B \mid A] \, \mathbb{P}[A]}{\mathbb{P}[B \mid A] \, \mathbb{P}[A] + \mathbb{P}[B \mid A^c] \, \mathbb{P}[A^c]} \end{aligned}$$

Random Variables

- **Random Variable**: A random variable is a real-valued outcome; a function from the sample space S to real numbers $\mathbb{R}$



- Example: Coin flip, we often use

$$X = \begin{cases} 1 & \text{if } H \\ 0 & \text{if } T \end{cases}$$

Random Variables

- X denote random variable
- $X = x$ denote X has a particular realization x
- **Support** of X: The set that random variable X is defined, denoted $\mathscr{X}$
- **Discrete** Random Variable: Sample space / support of X is finite or countable
- **Continuous** Random Variable: Sample space / support of X is uncountable

Distribution Function

- We can then associate random variables with probability!
- **Cumulative Distribution Function (CDF)** of a random variable X is the probability that X is less than or equal to some value x:

$$F_X(x) = \mathbb{P}[X \leq x]$$

- A CDF $F(x)$ must satisfy the following conditions:
  1. $F(x)$ is non-decreasing in x (because we're including more outcomes)
  2. $\lim_{x \to -\infty} F(x) = 0$ (probability of empty set)
  3. $\lim_{x \to \infty} F(x) = 1$ (probability of whole sample space)
  4. $F(x)$ is right-continuous (right limit must exist) (technical)
- Continuous random variable $\Leftrightarrow$ CDF is a continuous function
- Discrete random variable $\Leftrightarrow$ CDF is a step function

Density and Mass Functions

- For **continuous** random variable, its **Probability Density Function (PDF)** is

$$f(x) = F'_X(x) = \frac{d}{dx} F_X(x)$$

  - Fundamental Theorem of Calculus says $\mathbb{P}[a \le X \le b] = \int_a^b f(x)\,dx = F(b) - F(a)$
- For **discrete** random variable, its **Probability Mass Function (PMF)** is

$$f(x) = \mathbb{P}[X = x] \qquad [= F_X(x) - F_X(x - \varepsilon)]$$

  - Since discrete CDF is a step function, it is not differentiable everywhere
  - But we can still calculate that $\mathbb{P}[a \le X \le b] = \sum_{x \in \{a, \cdots, b\}} f(x) = \sum_{x \in \{a, \cdots, b\}} \mathbb{P}[X = x]$
- Either way, PDF and PMF must satisfy the following conditions:

1. $f(x) \ge 0$ for all x (positivity)
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (PDF) or $\sum_x f(x) = 1$ (PMF) (sums to 1)

Example: Bernoulli Random Variable

- $X \sim \text{Bernoulli}(p)$ if X has PMF

$$\mathbb{P}[X = 1] = p$$
$$\mathbb{P}[X = 0] = 1 - p$$

- The support of X is $\{0, 1\}$
- Note that we can also write the PMF as

$$\mathbb{P}[X = x \mid p] = p^x (1 - p)^{1-x} \mathbb{1}(x \in \{0, 1\})$$

- What is the CDF?

Example: Uniform Distribution

- $X \sim \mathrm{Uniform}[0, 1]$ if X has PDF

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- The support of X is $[0, 1]$
- We can also write the PDF as

$$f(x) = \mathbb{1}(x \in [0, 1])$$

- What is the CDF?

Expectation

- We often want to summarize characteristics of a distribution of a random variable
- What about we take the average of a random variable, weighted by probability?
- **Expectation**: The expected value of a random variable X is

$$\mathbb{E}[X] = \begin{cases} \int_{-\infty}^{\infty} x f(x)\, dx = \int_{-\infty}^{\infty} x\, dF(x) & \text{if } X \text{ continuous} \\ \sum_{x \in \mathcal{X}} x f(x) = \sum_{x \in \mathcal{X}} x\, \mathbb{P}[X = x] & \text{if } X \text{ discrete} \end{cases}$$

- Expectation tells us about the **central tendency** of a distribution
- More generally, we can find the expectation of any function of X

$$\mathbb{E}[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f(x)\, dx = \int_{-\infty}^{\infty} g(x)\, dF(x) & \text{if } X \text{ continuous} \\ \sum_{x \in \mathcal{X}} g(x) f(x) = \sum_{x \in \mathcal{X}} g(x)\, \mathbb{P}[X = x] & \text{if } X \text{ discrete} \end{cases}$$

Example

- Recall the Indicator Function: $\mathbb{1}(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$
- Show that $\mathbb{E}[\mathbb{1}(X \in A)] = \mathbb{P}[A]$

Example: Bernoulli Random Variable

- $X \sim \text{Bernoulli}(p)$ if X has PMF

$$\mathbb{P}[X = 1] = p$$
$$\mathbb{P}[X = 0] = 1 - p$$

- What is $\mathbb{E}[X]$?

- What is $\mathbb{E}[X^2]$?

Example: Uniform Distribution

- $X \sim \text{Uniform}[0, 1]$ if X has PDF

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- What is $\mathbb{E}[X]$?

- What is $\mathbb{E}[X^2]$?

Properties of Expected Values

1. Expectation of a constant is a constant

$$\mathbb{E}[c] = c$$

2. Constants come out

$$\mathbb{E}[cg(Y)] = c\,\mathbb{E}[g(Y)]$$

3. Expectation is **linear**: For any random variables $Y_1, \dots, Y_n$ (either dependent or independent),

$$\mathbb{E}[g(Y_1) + \cdots + g(Y_n)] = \mathbb{E}[g(Y_1)] + \cdots + \mathbb{E}[g(Y_n)]$$

4. If X and Y are **independent**, then the product is easy

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

5. Expected Value of Expected Values:

$$\mathbb{E}[\mathbb{E}[Y]] = \mathbb{E}[Y]$$

(since $\mathbb{E}[Y]$ is a constant)

Variance and Covariance

- How do we measure the distance of a random variable X from its mean?

$$X - \mathbb{E}[X]$$

- But we would like the "distance" to stay positive…
  - Absolute value $\rightsquigarrow$ Hard to deal with; not differentiable
  - Square $\rightsquigarrow$ Smoother, differentiable
- **Variance** tells us about the **spread** of the distribution around the center

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- **Covariance** measures the **co-movement** of two random variables around their own centers

$$\mathbb{C}\mathrm{ov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

- Normalize: $\mathrm{SD}[X] = \sqrt{\mathbb{V}\mathrm{ar}[X]}$, $\mathbb{C}\mathrm{orr}[X, Y] = \mathbb{C}\mathrm{ov}[X, Y] / \sqrt{\mathbb{V}\mathrm{ar}[X]\,\mathbb{V}\mathrm{ar}[Y]}$

Properties of Variance and Covariance

1. $\mathbb{V}\mathrm{ar}[c] = 0$
2. $\mathbb{V}\mathrm{ar}[a + bX] = b^2\, \mathbb{V}\mathrm{ar}[X]$
3. $\mathbb{C}\mathrm{ov}[a + bX, c + dY] = bd\, \mathbb{C}\mathrm{ov}[X, Y]$
4. $\mathbb{C}\mathrm{ov}[X + Z, Y + W] = \mathbb{C}\mathrm{ov}[X, Y] + \mathbb{C}\mathrm{ov}[X, W] + \mathbb{C}\mathrm{ov}[Z, Y] + \mathbb{C}\mathrm{ov}[Z, W]$
5. $\mathbb{V}\mathrm{ar}[X + Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\, \mathbb{C}\mathrm{ov}[X, Y]$

Example: Bernoulli Random Variable

- $X \sim \mathrm{Bernoulli}(p)$ if X has PMF

$$\mathbb{P}[X = 1] = p$$
$$\mathbb{P}[X = 0] = 1 - p$$

- What is $\mathbb{V}\mathrm{ar}[X]$?

Example: Uniform Distribution

- $X \sim \text{Uniform}[0, 1]$ if X has PDF

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- What is $\mathbb{V}\text{ar}[X]$?

Binomial Distribution

- Suppose we repeat the Bernoulli trial for n times
- Each trial i follows the same distribution

$$X_i \sim \text{Bernoulli}(p), \quad \text{so } \mathbb{P}[X_i = 1] = p \text{ for all } i$$

- Each trial also independent of each other

$$\mathbb{P}[X_i = x_i, X_j = x_j] = \mathbb{P}[X_i = x_i]\,\mathbb{P}[X_j = x_j]$$

- We want to count the number of successes, denoted by $Y = X_1 + \cdots + X_n$
- **One specific way** to obtain y number of success is:

$$\mathbb{P}[X_1 = 1, X_2 = 0, X_3 = 0, \ldots, X_n = 0] = \mathbb{P}[X_1 = 1]\,\mathbb{P}[X_2 = 0] \ldots \mathbb{P}[X_n = 0]$$

$$= \underbrace{\mathbb{P}[X_1 = 1]\,\mathbb{P}[X_3 = 1] \ldots}_{y \text{ terms}} \times \underbrace{\mathbb{P}[X_2 = 0] \ldots \mathbb{P}[X_n = 0]}_{n-y \text{ terms}}$$

$$= \underbrace{pp \cdots p}_{y \text{ terms}} \times \underbrace{(1-p) \cdots (1-p)}_{n-y \text{ terms}} = p^y (1-p)^{n-y}$$

Binomial Distribution

- This is just one instance of y number of successes
- There are $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ instances that we can get y successes
- So we have the PMF of the Binomial Distribution

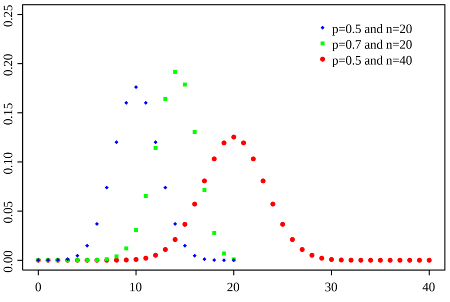
$$\mathbb{P}[Y = y \,|\, n, p] = \binom{n}{y} p^y (1-p)^{n-y} \mathbb{1}(y \in \{0, 1, \ldots, n\})$$

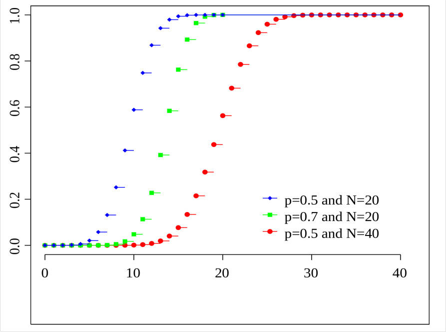
- The support of Y is $\{0, 1, \ldots, n\}$
- This example shows that:

$$\text{If} \quad X_i \sim \textbf{i.i.d.}\ \text{Bernoulli}(p) \quad \Rightarrow \quad X_1 + \cdots + X_n = Y \sim \text{Binomial}(n, p)$$

  - Note that the i.i.d. (independent and identically distributed) assumption is crucial in our derivation

Binomial PMF and CDF





Binomial Distribution

- $X_i \sim$ iid Bernoulli(p), $X_1 + \cdots + X_n = Y \sim$ Binomial (n, p)
- What is $\mathbb{E}[Y]$?

- What is $\mathbb{V}\mathrm{ar}[Y]$?

Poisson Distribution

- **Poisson distribution** is often used to model **rare event counts**
  - Counts of the number of events that occur during some unit of time
  - The event would occur with a fixed "arrival rate" $\lambda > 0$
  - ex. Number of wars in a year (assuming "arrival rate" is fixed)
- $X \sim \text{Poisson}(\lambda)$ if X has PMF

$$\mathbb{P}[X = x \mid \lambda] = \frac{e^{-\lambda}\lambda^x}{x!}\mathbb{1}(x \in \{0, 1, 2, \ldots\})$$

- The support of X is $\{0, 1, 2, \ldots\}$

Poisson Distribution

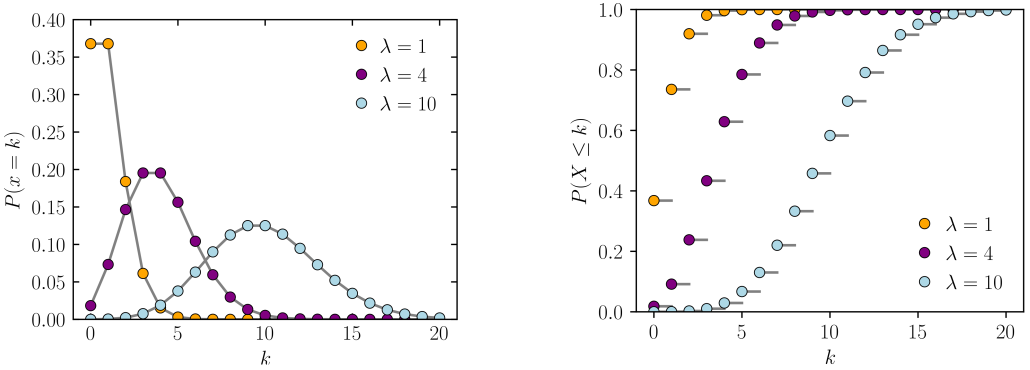
- $X \sim \text{Poisson}(\lambda)$ if X has PMF

$$\mathbb{P}[X = x \mid \lambda] = \frac{e^{-\lambda}\lambda^{x}}{x!}\mathbb{1}(x \in \{0, 1, 2, \ldots\})$$

- Is this a legitimate PMF? (Does it follow the two conditions?)
  - Note the Taylor expansion of e^x around $x = 0$:

$$e^{x} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Poisson PMF and CDF



Poisson Distribution

- $X \sim \text{Poisson}(\lambda)$ if X has PMF

$$\mathbb{P}[X = x \,|\, \lambda] = \frac{e^{-\lambda}\lambda^x}{x!}\mathbb{1}(x \in \{0, 1, 2, \ldots\})$$

- What is $\mathbb{E}[X]$?
  - Hint: Use the fact that PMF sums to 1: $\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = 1$

Poisson Distribution

- $X \sim \text{Poisson}(\lambda)$ if X has PMF

$$\mathbb{P}[X = x \mid \lambda] = \frac{e^{-\lambda}\lambda^x}{x!}\mathbb{1}(x \in \{0, 1, 2, \ldots\})$$

- Show that $\mathbb{E}[X^2] = \lambda(\lambda + 1)$
  - Hint: Use the fact that PMF sums to 1: $\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = 1$

- Show that $\mathbb{V}\text{ar}[X] = \lambda$