Seminar 1: Discrete Random Variables

Carlos Buitrago

Table of contents

1 Probability Distributions and Random Variables

In the previous lectures, probabilities were assigned directly to events in a probability space $(\Omega, \mathcal{F}, P)$. In practice, however, we are often interested in numerical quantities associated with the outcomes of an experiment, such as the number of heads obtained when tossing a coin several times, the number of customers arriving at a store during one hour, or the height of a randomly selected individual.

Such quantities are modeled by **random variables**. To study them, we first introduce probability distributions on the real line.

1.1 Borel σ-Algebra

The **Borel σ-algebra** on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra containing all open subsets of $\mathbb{R}$.

Elements of $\mathcal{B}(\mathbb{R})$ are called **Borel sets**. In particular, every open interval, closed interval, half-open interval, countable union, countable intersection, and complement of such sets belongs to $\mathcal{B}(\mathbb{R})$.

1.2 Measurable Functions and Random Variables

Let $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be measurable spaces. A function

$$\xi : \Omega \to \mathbb{R}$$

is called **measurable** if

$$\xi^{-1}(B) = \{\omega \in \Omega : \xi(\omega) \in B\} \in \mathcal{F} \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

A **random variable** on a probability space $(\Omega, \mathcal{F}, P)$ is simply a measurable function $\xi : \Omega \to \mathbb{R}$.

We denote by

$$X_\xi = \{\xi(\omega) : \omega \in \Omega\}$$

the set of all possible values of ξ.

If X_ξ is finite or countable, then ξ is called a **discrete random variable**.

1.3 Probability Distributions

A **probability distribution** is any probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

That is,

$$P : \mathcal{B}(\mathbb{R}) \to [0, 1].$$

If there exists a finite or countable set $X \subseteq \mathbb{R}$ such that $P(X) = 1$, then P is called a **discrete probability distribution**.

1.4 Distribution of a Random Variable

Let ξ be a random variable on $(\Omega, \mathcal{F}, P)$.

The **probability distribution of** ξ is the probability measure P_ξ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$P_\xi(B) = P(\{\omega \in \Omega : \xi(\omega) \in B\}) = P(\xi \in B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Thus, the distribution of a random variable completely describes the probabilities of all events involving its values.

1.5 Probability Mass Function (PMF)

Let ξ be a discrete random variable and let $X_\xi = \{x_1, x_2, ...\}$.

The **probability mass function** (PMF) of ξ is the function

$$p_\xi : \mathbb{R} \to [0, 1]$$

defined by

$$p_\xi(x_k) = P(\xi = x_k), \qquad x_k \in X_\xi,$$

and $ p_□(x)=0$, $x \notin X_\xi$

The PMF completely determines the distribution of a discrete random variable and satisfies $\sum_{x \in X_\xi} p_\xi(x) = 1$.

2 Examples of Discrete Probability Distributions

2.1 Uniform Distribution

Let X be a finite set. The uniform distribution on X is defined by

$$P(x) = \frac{1}{|X|}, \qquad x \in X.$$

2.2 Bernoulli Distribution

For $p \in (0, 1)$, $X = \{0, 1\}$, and

$$P(0) = 1 - p, \qquad P(1) = p.$$

We write

$$\xi \sim \text{Ber}(p).$$

2.3 Binomial Distribution

For $n \in \mathbb{N}$ and $p \in (0, 1)$, $X = \{0, 1, \dots, n\}$, and

$$P(k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

We write

$$\xi \sim \text{Bin}(n, p).$$

2.4 Geometric Distribution

For $p \in (0,1)$, $X = \mathbb{N} \setminus \{0\}$, and

$$P(k) = (1-p)^{k-1}p, \qquad k \geq 1.$$

We write

$$\xi \sim \mathrm{Geom}(p).$$

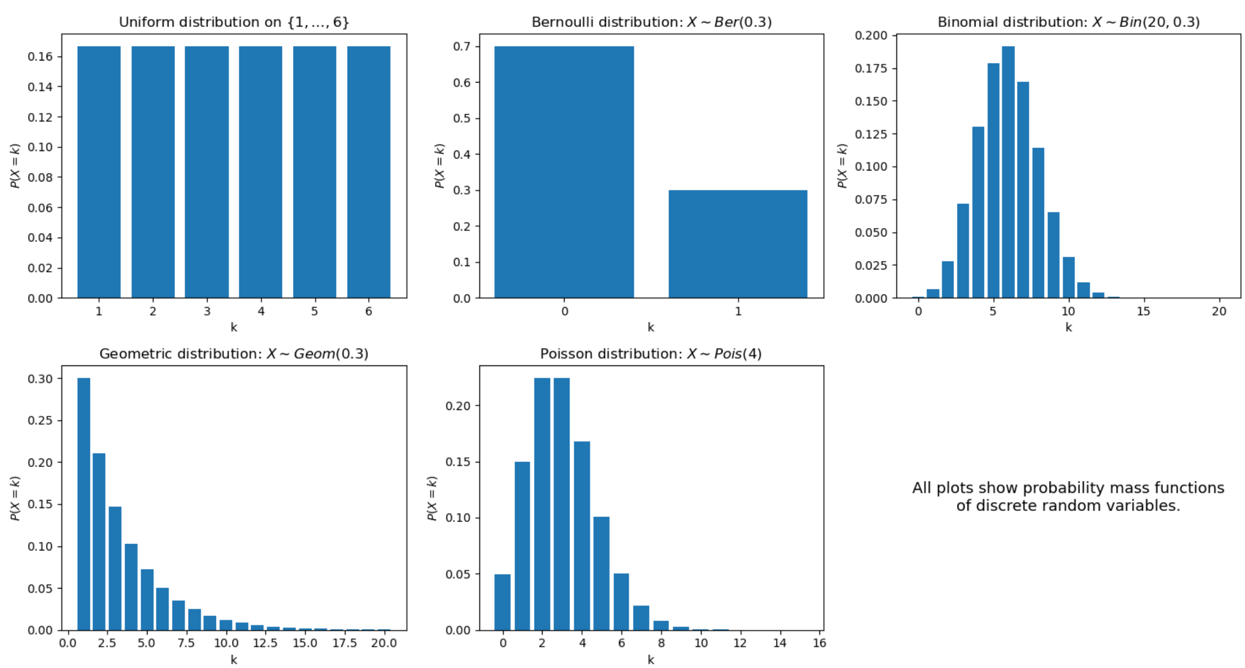
2.5 Poisson Distribution

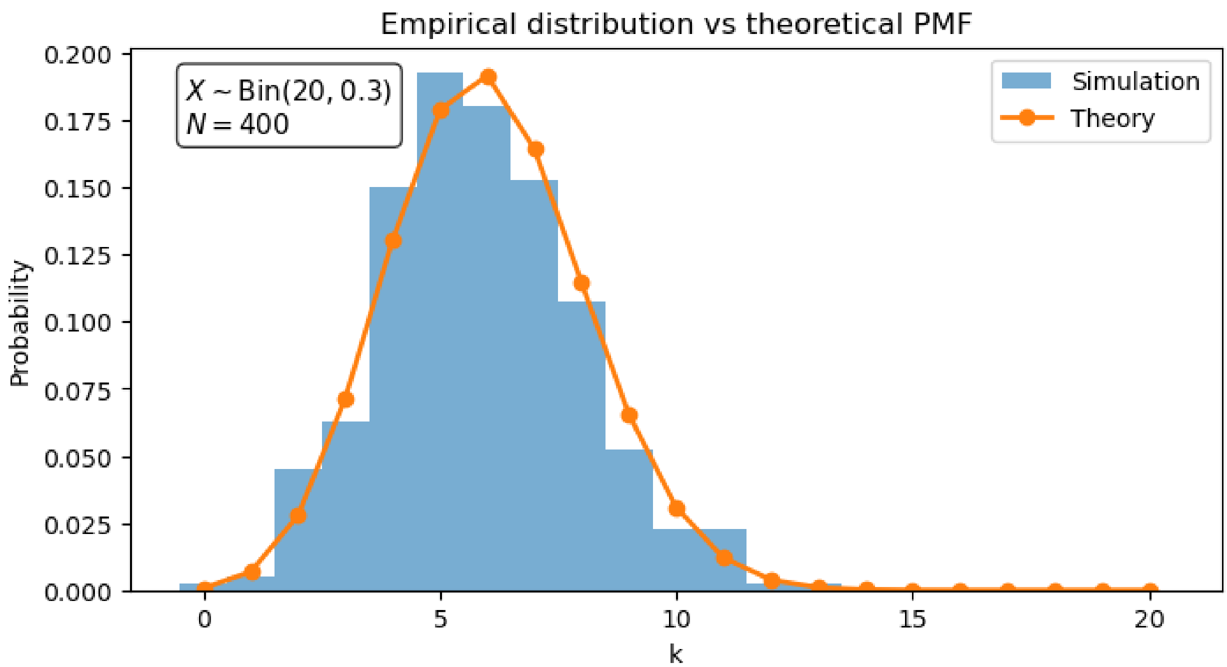
For $\lambda > 0$, $X = \{0, 1, 2, \ldots\}$, and

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

We write

$$\xi \sim \mathrm{Pois}(\lambda).$$





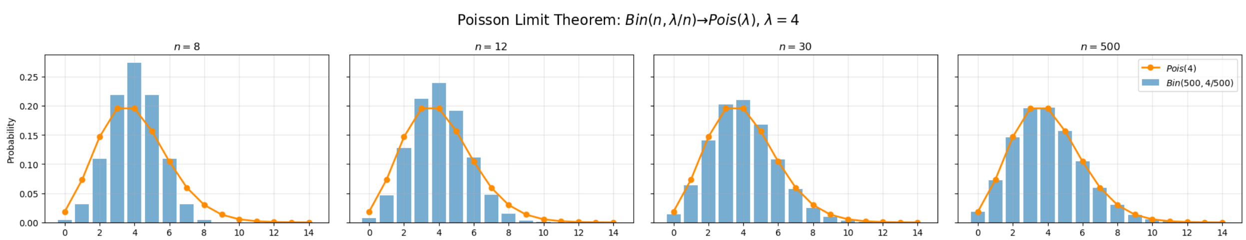
3 Poisson Limit Theorem

Let $\xi_n \sim \mathrm{Bin}(n, p_n)$ and suppose that $np_n \to \lambda > 0$.

Then for every fixed $k \geq 0$,

$$P(\xi_n = k) \longrightarrow \frac{\lambda^k e^{-\lambda}}{k!}.$$

In other words, the binomial distribution converges to the Poisson distribution when the number of trials becomes large and the success probability becomes small in such a way that the expected number of successes remains approximately constant.



4 Independence of Random Variables

Two discrete random variables ξ and η are called **independent** if

$$P(\xi = x, \eta = y) = P(\xi = x)P(\eta = y)$$

for all possible values $x \in X_\xi$ and $y \in X_\eta$.

Similarly, a collection of random variables $\xi_1, \dots, \xi_n$ is called mutually independent if

$$P(\xi_{i_1} = x_{i_1}, \dots, \xi_{i_k} = x_{i_k}) = \prod_{j=1}^{k} P(\xi_{i_j} = x_{i_j})$$

for every subset $\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$.

4.1 Convolution of Discrete Distributions

Let ξ and η be independent discrete random variables. Then

$$P(\xi + \eta = k) = \sum_{i=-\infty}^{+\infty} P(\xi = i) P(\eta = k - i).$$

This formula is called the **convolution formula**.

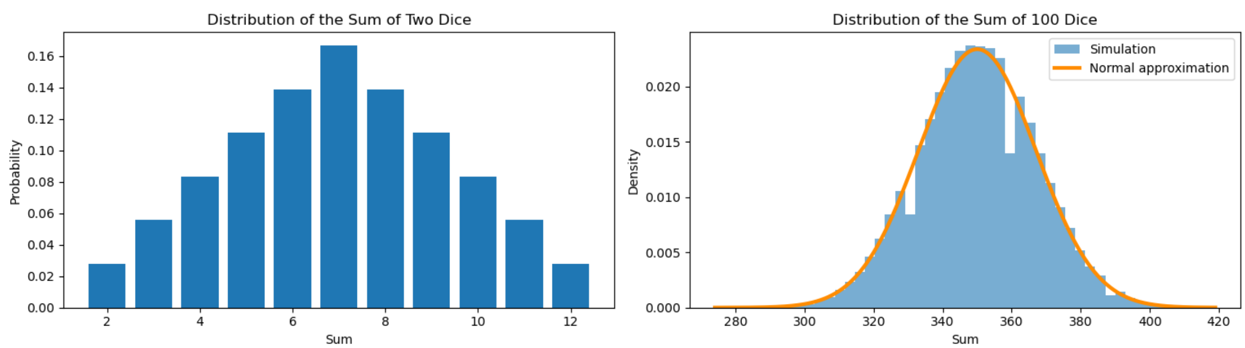
Important examples:

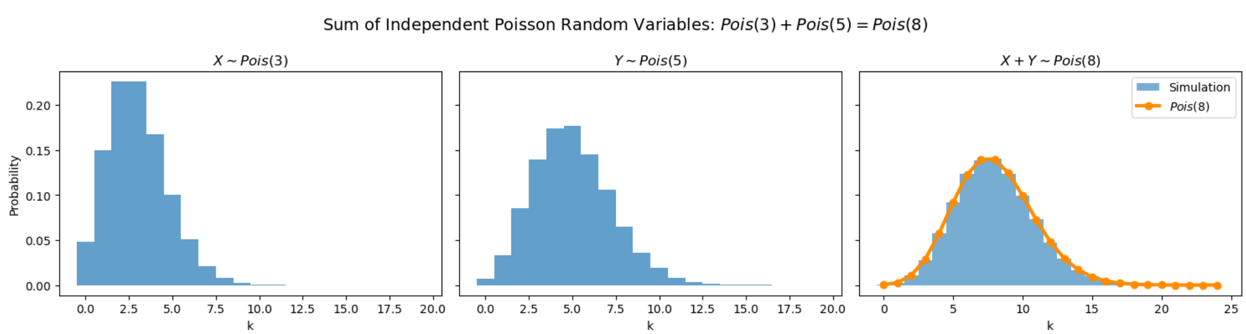
- If $\xi \sim \mathrm{Bin}(n, p)$ and $\eta \sim \mathrm{Bin}(m, p)$ are independent, then

$$\xi + \eta \sim \mathrm{Bin}(n + m, p).$$

- If $\xi \sim \mathrm{Pois}(\lambda_1)$ and $\eta \sim \mathrm{Pois}(\lambda_2)$ are independent, then

$$\xi + \eta \sim \mathrm{Pois}(\lambda_1 + \lambda_2).$$





5 Discrete Random Vectors

So far we have studied a single random variable $\xi : \Omega \to \mathbb{R}$. In many applications, however, several random quantities are observed simultaneously. This leads to the notion of a random vector.

5.1 Definition

A **2-dimensional discrete random vector** is a pair of discrete random variables

$$(\xi, \eta).$$

More generally,

$$(\xi_1, \xi_2, \dots, \xi_n).$$

For every outcome $\omega \in \Omega$, the random vector assigns a point

$$(\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega)) \in \mathbb{R}^n.$$

The set of all possible values is called the **support** of the random vector.

5.2 Joint Probability Mass Function

The probability distribution of a discrete random vector is described by its **joint probability mass function (joint PMF)**.

5.2.1 Definition

For a discrete random vector (ξ, η),

$$p_{\xi,\eta}(x, y) = \mathbb{P}(\xi = x, \eta = y).$$

The value $p_{\xi,\eta}(x, y)$ is the probability that both events occur simultaneously.

The joint PMF satisfies $p_{\xi,\eta}(x, y) \geq 0$, and

$$\sum_x \sum_y p_{\xi,\eta}(x, y) = 1.$$

5.3 Joint PMF Tables

A convenient way to represent a joint PMF is by a table.

5.3.1 Example: Two Fair Dice

For two independent fair dice,

$$p_{\xi,\eta}(x, y) = \frac{1}{36}, \qquad x, y \in \{1, \dots, 6\}.$$

$\xi \backslash \eta$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Every cell corresponds to one outcome (x, y).

The sum of all entries of the table must be equal to 1.

6 Marginal Distributions

The distributions of ξ and η can be recovered from the joint PMF.

6.1 Definition

The **marginal PMF** of ξ is

$$p_\xi(x) = \sum_y p_{\xi,\eta}(x, y).$$

Similarly,

$$p_\eta(y) = \sum_x p_{\xi,\eta}(x, y).$$

In words, we sum over all possible values of the other variable.

6.1.1 Example

For two fair dice,

$$p_\xi(x) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6}.$$

6.2 Independence

The notion of independence extends naturally to random vectors.

6.2.1 Definition

Random variables ξ and η are **independent** if

$$p_{\xi,\eta}(x, y) = p_\xi(x)p_\eta(y) \qquad \forall x, y.$$

Thus, the probability of a pair of outcomes factors into the product of the marginal probabilities.

7 Discrete distributions in ML

7.1 Logistic Regression

Many machine learning problems involve predicting a binary outcome.

Suppose

$$Y = \begin{cases} 1, & \text{success}, \\ 0, & \text{failure}. \end{cases}$$

Then

$$Y \sim Ber(p),$$

where p is the probability of success. Logistic regression is used when the response variable has only two possible outcomes.

Examples include:

- Will a customer purchase a product?
- Will a student pass an exam?
- Will a patient develop a disease?
- Is an email spam or not spam?
- Will a user click on an advertisement?

In logistic regression, this probability depends on explanatory variables. For example, an online store may use

$$(x_1, x_2, x_3) = (\text{age, income, number of previous purchases})$$

to predict the probability that a customer makes a purchase.

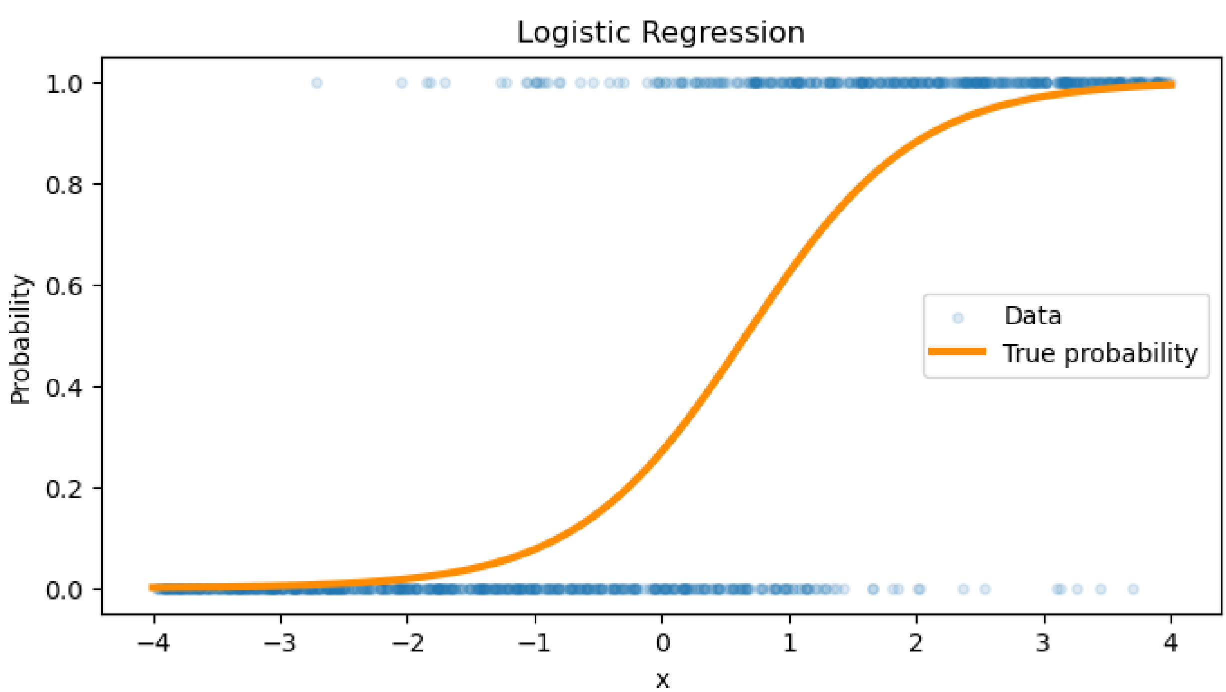
For a single feature x, the model assumes

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Thus,

$$Y \mid x \sim Ber(p(x)).$$

Logistic regression is one of the most widely used classification methods in machine learning.



7.2 Poisson Regression

Many applications involve predicting counts.

Examples include:

- number of website visits;
- number of accidents;
- number of insurance claims;
- number of emails received.

In such situations, it is natural to assume

$$Y \sim Pois(\lambda),$$

where λ is the expected count.

Poisson regression models the parameter λ as

$$\lambda(x) = e^{\beta_0 + \beta_1 x}.$$

The exponential function guarantees that $\lambda(x) > 0$.

For example, a company may use

$$(x_1, x_2, x_3) = (\text{advertising budget, season, number of users})$$

to predict the expected number of website visits. Thus,

$$Y \mid x \sim Pois(\lambda(x)).$$

Poisson regression is one of the most important generalized linear models.