Key Distributions - Fall 2024

Instructor: Dr. Juhee Lee
TA: Antonio Aguirre

University of California, Santa Cruz
Fall 2024

Introduction

This document provides a comprehensive overview of six key probability distributions that are widely used across the sciences, engineering, and industry. These distributions are mathematical models designed to describe and predict the behavior of random quantities that vary in uncertain ways.

Understanding probability distributions is crucial because they bridge theoretical statistics with real-world applications. Each distribution in this document represents a unique type of **random behavior**.

For many students, one common challenge is recognizing that a probability distribution is not just a formula—it is a **model**. A model is an abstraction of reality, built to simplify and represent complex processes. For example:

- A Binomial distribution models the number of successes in repeated independent trials, such as the number of customers who make a purchase.
- A Normal distribution models continuous data that clusters around an average, such as heights in a population or errors in measurements.

This document emphasizes the random quantity being modeled to help you connect mathematical theory with practical contexts. Remember, the goal of probability distributions is not just computation but also **understanding how to describe and predict randomness** in meaningful ways.

As you study these models, keep in mind:

- The formulas are tools, not just abstract math—they are designed to capture real patterns in data.
- Each distribution has specific assumptions. These assumptions matter; they ensure that the model fits the problem.
- Probability distributions are versatile and foundational, forming the basis for countless methods in data science, machine learning, economics, biology, and more.

1. Bernoulli Distribution

Definition

The Bernoulli distribution models the outcome of a single trial with two possible outcomes: success (1) or failure (0). The probability mass function is:

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}.$$

Parameters

p**:** Probability of success ($0 \leq p \leq 1$).

Support

$$x \in \{0, 1\}$$

Assumptions

- A single trial with two possible outcomes: success (1) or failure (0).
- The probability of success (p) is fixed.

Mean and Variance

Mean: $\mathbb{E}[X] = p$

Variance: $\text{Var}(X) = p(1-p)$

Application

In quality control, the Bernoulli distribution is used to model whether a product passes (1) or fails (0) inspection.

Random Quantity

The Bernoulli distribution models the outcome of a single trial.

2. Binomial Distribution

Definition

The Binomial distribution models the number of successes in n independent trials, each with a success probability p. The probability mass function is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

Parameters

n: Number of independent trials.

p: Probability of success in each trial.

Support

$$x \in \{0, 1, \ldots, n\}$$

Assumptions

- Trials are independent.
- The probability of success p is constant across all trials.
- Sampling is either with replacement or from a large population.

Mean and Variance

Mean: $\mathbb{E}[X] = np$

Variance: $\mathrm{Var}(X) = np(1-p)$

Application

In surveys, the Binomial distribution is used to model the number of respondents who answer positively out of n total participants.

Random Quantity

The Binomial distribution models the number of successes in n independent trials.

3. Hypergeometric Distribution

Definition

The Hypergeometric distribution models the number of successes in n dependent draws from a finite population containing K successes. The probability mass function is:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k \in \{\max(0, n + K - N), \ldots, \min(n, K)\}.$$

Parameters

N**:** Total population size.

K**:** Number of successes in the population.

n**:** Sample size.

Support

$$x \in \{\max(0, n + K - N), \ldots, \min(n, K)\}$$

Assumptions

- Sampling is without replacement.
- Trials are dependent because the population is finite.

Mean and Variance

Mean: $\mathbb{E}[X] = n\frac{K}{N}$

Variance: $\text{Var}(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

Application

In wildlife studies, the Hypergeometric distribution models the number of tagged animals recaptured in a sample.

Random Quantity

The Hypergeometric distribution models the number of successes in n dependent draws from a finite population.

4. Negative Binomial Distribution

Definition

The Negative Binomial distribution models the number of failures before achieving r successes in independent Bernoulli trials. The probability mass function is:

$$P(X = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \quad k \in \{0, 1, 2, \dots\}.$$

Parameters

r: Number of successes required.

p: Probability of success in each trial.

Support

$$x \in \{0, 1, 2, \dots\}$$

Assumptions

- Trials are independent.
- The probability of success p is fixed.

Mean and Variance

Mean: $\mathbb{E}[X] = \frac{r(1-p)}{p}$

Variance: $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

Application

In sports, the Negative Binomial distribution models the number of missed shots before scoring r successful goals.

Random Quantity

The Negative Binomial distribution models the number of failures before achieving r successes.

5. Poisson Distribution

Definition

The Poisson distribution models the number of events occurring in a fixed interval of time or space, assuming events happen independently and at a constant rate λ. The probability mass function is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \{0, 1, 2, \dots\}.$$

Parameters

λ: Average number of events per fixed interval ($\lambda > 0$).

Support

$$x \in \{0, 1, 2, \dots\}$$

Assumptions

- Events occur independently of one another.
- The rate λ is constant over the fixed interval.

Mean and Variance

Mean: $\mathbb{E}[X] = \lambda$

Variance: $\text{Var}(X) = \lambda$

Application

In traffic analysis, the Poisson distribution is used to model the number of cars passing through a checkpoint in an hour.

Random Quantity

The Poisson distribution models the number of events occurring in a fixed interval of time or space.

6. Normal (Gaussian) Distribution

Definition

The Normal distribution models continuous data that clusters symmetrically around a mean μ, with a bell-shaped curve. The probability density function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Parameters

μ: Mean (center of the distribution).

σ^2: Variance (spread of the distribution, $\sigma^2 > 0$).

Support

$$x \in \mathbb{R}$$

Assumptions

- Data is symmetrically distributed around the mean μ.
- Deviations from the mean follow an exponential decay, with larger deviations being less likely.

Mean and Variance

Mean: $\mathbb{E}[X] = \mu$

Variance: $\text{Var}(X) = \sigma^2$

Application

In biology, the Normal distribution is commonly used to model the distribution of human heights within a population.

Random Quantity

The Normal distribution models the value of a continuous variable.

Exercises

Problem 1: Teaching Success

Antonio, a dedicated TA, gives a weekly statistics quiz to his class of 10 students. Historically, Antonio has observed that each student has a 70% chance of passing the quiz. Assume that the outcomes for each student are independent.

(a) What is the probability that exactly 7 students pass the quiz this week?

(b) What is the probability that at least 8 students pass the quiz?

(c) What is the expected number of students who will pass the quiz, and what is the variance?

Problem 2: Emails Before the Final

The day before the final exam, Antonio receives an overwhelming number of emails from students asking for full explanations of specific topics they haven't studied. Historically, Antonio receives an average of 5 such emails per hour the day before the final. Assume these emails arrive according to a Poisson process.

(a) What is the probability that Antonio receives exactly 3 emails in a one-hour period?

(b) What is the probability that Antonio receives more than 7 emails in a one-hour period?

(c) If Antonio checks his email over a 2-hour period, what is the expected number of emails, and what is the standard deviation?

Problem 3: Grading Speeds

Despite his underpayment as a TA, Antonio grades exams at an average speed of 12 minutes per exam with a standard deviation of 3 minutes. Note: his pay rate has nothing to do with the problem.

(a) What is the probability that Antonio grades an exam in less than 10 minutes?

(b) What is the probability that Antonio takes between 11 and 14 minutes to grade an exam?

(c) What is the probability that Antonio takes between 6 and 18 minutes to grade an exam?

(d) What is the probability that Antonio grades an exam in more than 21 minutes?