Statistics & Probability

Chapter 3: DISCRETE RANDOM VARIABLE AND PROBABILITY DISTRIBUTIONS

FPT University

Department of Mathematics

Quy Nhon, 2023

Table of Contents

Table of Contents

Discrete Random Variables

Discrete Random Variable

A **discrete random variable** is a random variable with a finite (or countably infinite) range.

Example (Discrete Random Variable)

1. Roll a fair die twice. Let X be the number of times 4 comes up. Then $x = 0, 1$, or 2.
2. Toss a fair coin 5 times. Let X be the number of heads. Then $x = 0, 1, 2, 3, 4$, or 5.
3. X is the number of stocks in the Dow Jones Industrial Average that have share price increases on a given day, then X is a discrete random variable because whose share price increases can be counted.

Table of Contents

Let X be a discrete random variable with possible outcomes $x_1, x_2, \ldots, x_n$.

1. Find the probability of each possible outcomes.
2. Check that each probability is between 0 and 1, and the sum is 1.
3. Summarizing results in following table, we obtain the probability distribution of X.

X	x_1	x_2	$\ldots$	x_n
$P(x)$	p_1	p_2	$\ldots$	p_n

Example

Let the random variable X denote the number of heads in three tosses of a fair coin. Determine the probability distribution of X.

Solution. Sample space: $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

The events

- $[X = 0] = \{TTT\}$.
- $[X = 1] = \{HTT, THT, TTH\}$.
- $[X = 2] = \{HHT, HTH, THH\}$.
- $[X = 3] = \{HHH\}$.

X	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

Probability Mass Function

For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$, a **probability mass function** (PMF) is a function such that

1. $f(x_i) \geq 0$
2. $\sum_{i=1}^{n} f(x_i) = 1$
3. $f(x_i) = P(X = x_i)$.

Note

1. Sometimes PMF is also known as the discrete density function or frequency function.
2. The PMF is often the primary means of defining a discrete probability distribution.

Example (Probability Mass Function)

Suppose that a day's production of 850 manufactured parts contains 50 parts that do not conform to customer requirements. Two parts are selected at random, without replacement, from the batch. Let the random variable X equal the number of nonconforming parts in the sample. What is the probability mass function of X?

Solution. We have

$$P(X=0) = \frac{800}{850}\cdot\frac{799}{849} \approx 0.886$$
$$P(X=1) = 2\cdot\frac{800}{850}\cdot\frac{50}{849} \approx 0.111$$
$$P(X=2) = \frac{50}{850}\cdot\frac{49}{849} \approx 0.003.$$

Thus, the probability mass function of X is

$$f(0) = 0.886,\ f(1) = 0.111,\ f(2) = 0.003,\ f(x) = 0 \text{ otherwise}.$$

Table of Contents

Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a discrete random variable X, denoted as $F(x)$, is

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

Properties of Cumulative Distribution Functions

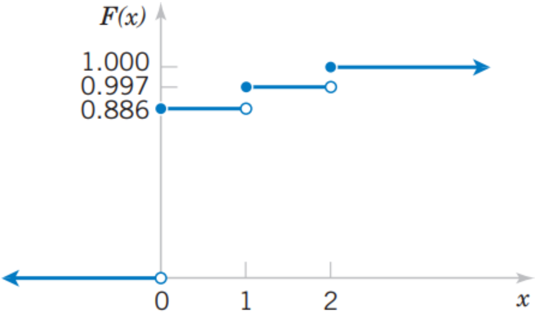
Let X be a discrete random variable with cumulative distribution function F. Then F satisfies the following:

1. $0 \leq F(x) \leq 1$.
2. If $x \leq y$ then $F(x) \leq F(y)$.
3. F is non-decreasing, i.e., F may be constant, but otherwise it is increasing.
4. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
5. $F(x)$ is right continuous at every x in $\mathbb{R}$.

Example (Cumulative Distribution Function)

X	0	1	2
$f(x)$	0.886	0.111	0.003

which implies $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.886 & \text{if } 0 \le x < 1 \\ 0.997 & \text{if } 1 \le x < 2 \\ 1 & \text{if } x \ge 2. \end{cases}$



Note

The function $F(x)$ is defined for all x from $-\infty < x < +\infty$ and not only for $0, 1$ and 2.

Let X be a discrete random variable that has the probability distribution table as follows

X	x_1	x_2	...	x_n
$P(x)$	p_1	p_2	...	p_n

where $x_1 < x_2 < \cdots < x_n$. Then the cumulative distribution function of X is

$$F(x) = \begin{cases} 0 & \text{if } x < x_1 \\ p_1 & \text{if } x_1 \le x < x_2 \\ p_1 + p_2 & \text{if } x_2 \le x < x_3 \\ \vdots & \\ p_1 + p_2 + \cdots + p_{n-1} & \text{if } x_{n-1} \le x < x_n \\ 1 & \text{if } x_n \le x. \end{cases}$$

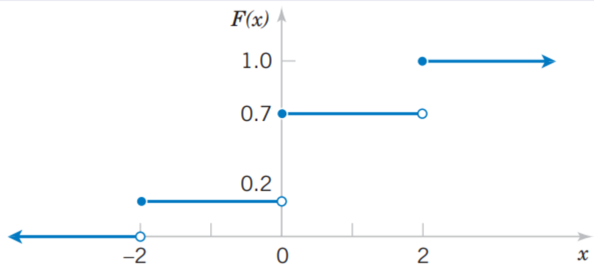
FUQN

Example (Cumulative Distribution Function)

Determine the probability mass function of X from the following cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < -2 \\ 0.2 & \text{if } -2 \le x < 0 \\ 0.7 & \text{if } 0 \le x < 2 \\ 1 & \text{if } 2 \le x. \end{cases}$$

Solution. The probability mass function at each point is the change in the cumulative distribution function at the point. Thus,



$$f(-2) = 0.2 - 0 = 0.2$$
$$f(0) = 0.7 - 0.2 = 0.5$$
$$f(2) = 1 - 0.7 = 0.3.$$

Quiz

Let $F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.5 & \text{if } 1 \leq x < 3 \\ 1 & \text{if } 3 \leq x \end{cases}$. Determine the following probabilities:

1. $P(X \leq 3)$.
2. $P(X \leq 2)$.
3. $P(1 \leq X \leq 2)$.
4. $P(X > 2)$.

FUQN

Table of Contents

Let X be a discrete random variable.

Mean

The **mean** or **expected value** of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x f(x).$$

Variance

The **variance** of X, denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2.$$

Standard deviation

The **standard deviation** of X is $\sigma = \sqrt{V(X)}$.

Properties of Mean

Let X, Y be any discrete random variables and a, b be arbitrary constants. Then

1. $E(a) = a$.
2. $E(aX) = aE(X)$.
3. $E(X+Y) = E(X) + E(Y)$.
4. $E(aX+b) = aE(X) + b$.
5. $E(XY) = E(X)E(Y)$ if X and Y are independent.

Properties of Variance

Let X, Y be any discrete random variables and a, b be arbitrary constants. Then

1. $V(a) = 0$.
2. $V(X) \geq 0$.
3. $V(aX) = a^2V(X)$.
4. $V(X+b) = V(X)$.
5. $V(aX+b) = a^2V(X)$.
6. $V(X) = E(X^2) - [E(X)]^2$.
7. $V(X+Y) = V(X) + V(Y)$ if X and Y are independent.

Example

The number of messages sent per hour over a computer network has the following distribution:

x	10	11	12	13	14	15
$f(x)$	0.08	0.15	0.30	0.20	0.20	0.07

Determine the mean and standard deviation of the number of messages sent per hour.

Solution. We have

$$E(X) = 10 \cdot 0.08 + 11 \cdot 0.15 + 12 \cdot 0.3 + 13 \cdot 0.2 + 14 \cdot 0.2 + 15 \cdot 0.07 = 12.5$$

$$V(X) = 10^2 \cdot 0.08 + 11^2 \cdot 0.15 + 12^2 \cdot 0.3 + 13^2 \cdot 0.2 + 14^2 \cdot 0.2 + 15^2 \cdot 0.07 - 12.5^2$$
$$= 1.85.$$
$$\sigma = \sqrt{V(X)} = \sqrt{1.85} = 1.36.$$

Table of Contents

Discrete Uniform Distribution

A random variable X has a **discrete uniform distribution** if each of the n values in its range, say, $x_1, x_2, \ldots, x_n$ has equal probability. Then, the PMF is given by

$$f(x_i) = P(X = x_i) = \frac{1}{n}.$$

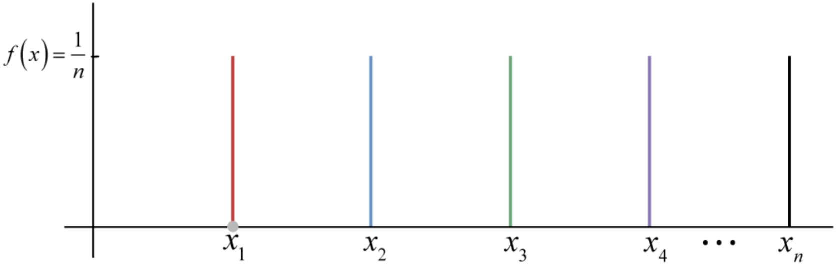


Figure. Discrete uniform distribution graph.

Remark

Let X be a discrete uniform random variable take the values $x = 1, 2, \dots, n$.

1. The probability mass function (PMF) is $f(x) = P(X = x) = \dfrac{1}{n}$.
2. The mean (expected value) is $E(X) = \dfrac{n+1}{2}$.
3. The variance is $V(X) = \dfrac{n^2 - 1}{12}$.

Generalization

Suppose X is a discrete uniform random variable on the consecutive integers $a, a+1, a+2, \ldots, b$ for $a \leq b$.

1. The probability mass function (PMF) of X is
$$f(x) = P(X = x) = \frac{1}{b-a+1}.$$
2. The mean (expected value) of X is
$$\mu = E(X) = \frac{b+a}{2}.$$
3. The variance of X is
$$\sigma^2 = V(X) = \frac{(b-a+1)^2 - 1}{12}.$$
4. The cumulative distribution function (CDF) of X is
$$F(x) = P(X \leq x) = \frac{x-a+1}{b-a+1} \text{ where } a \leq x \leq b.$$

Note. We can write $X \sim U(a, b)$.

Example

A box of 12 donuts sitting on the table, and you are asked to randomly select one donut without looking. Each of the 12 donuts has an equal chance of being selected. Identify the distribution, and calculate mean and variance.

Solution. The probability of any one donut being chosen is the same or uniform.

1. PMF: $f(x) = \dfrac{1}{n} = \dfrac{1}{12}$.
2. Mean: $E(X) = \dfrac{b+a}{2} = \dfrac{12+1}{2} = 6.5$.
3. Variance: $V(X) = \dfrac{(b-a+1)^2 - 1}{12} = \dfrac{(12-1+1)^2 - 1}{12} = \dfrac{143}{12}$.

Example

Roll a six-sided fair die. Suppose X denotes the number appear on the top of a die.

1. Find the probability that an even number appear on the top.
2. Find the probability that the number appear on the top is less than 3.
3. Compute mean and variance of X.

Solution. Let X denote the number appear on the top of a die. Then, the random variable X take the values $x = 1, 2, 3, 4, 5, 6$ and follows discrete uniform distribution. The PMF of X is

$$P(X = x) = \frac{1}{6 - 1 + 1} = \frac{1}{6}.$$

1. $P(X = \text{even number}) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5.$
2. $P(X < 3) = P(X = 1) + P(X = 2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$
3. $E(X) = \frac{6+1}{2} = 3.5$ and $V(X) = \frac{6^2 - 1}{12} = \frac{35}{12}.$

Quiz

A telephone number is selected at random from a directory. Suppose X denote the last digit of selected telephone number. Find the probability that the last digit of the selected number is

1. 5.
2. less than 5.
3. greater than or equal to 7.

Table of Contents

Binomial Distribution

A random experiment consists of n Bernoulli trials such that

1. the trials are independent,
2. each trial results in only two possible outcomes, labeled as "success" and "failure",
3. the probability of a success in each trial, denoted as p, remains constant.

The random variable X that equals the number of trials that result in a success has a **binomial random variable** with parameters $0 < p < 1$ and $n = 1, 2, \ldots$

The probability mass function (PMF) of X is

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $x = 0, 1, 2, \ldots, n$. We can write $X \sim B(n, p)$.

Mean and Variance

Let X be a binomial random variable with parameters p and n. Then

$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1-p).$$

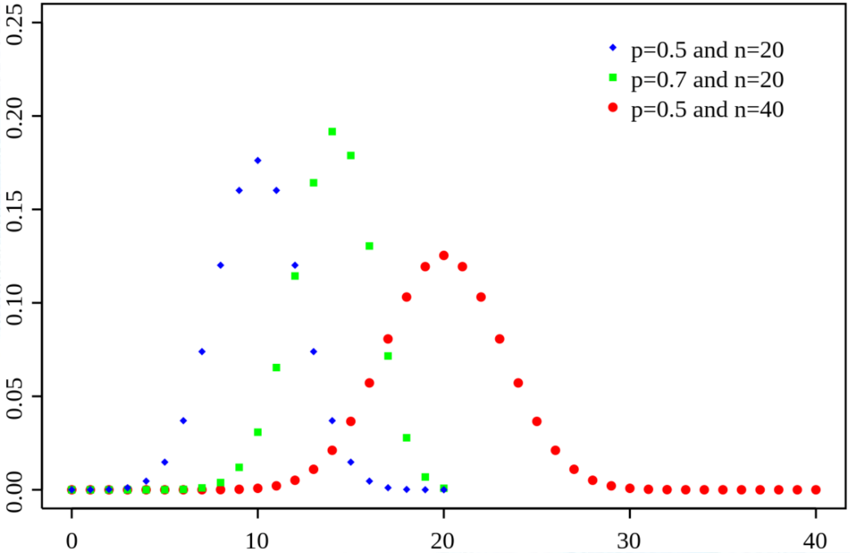


Figure. Probability mass function graph.

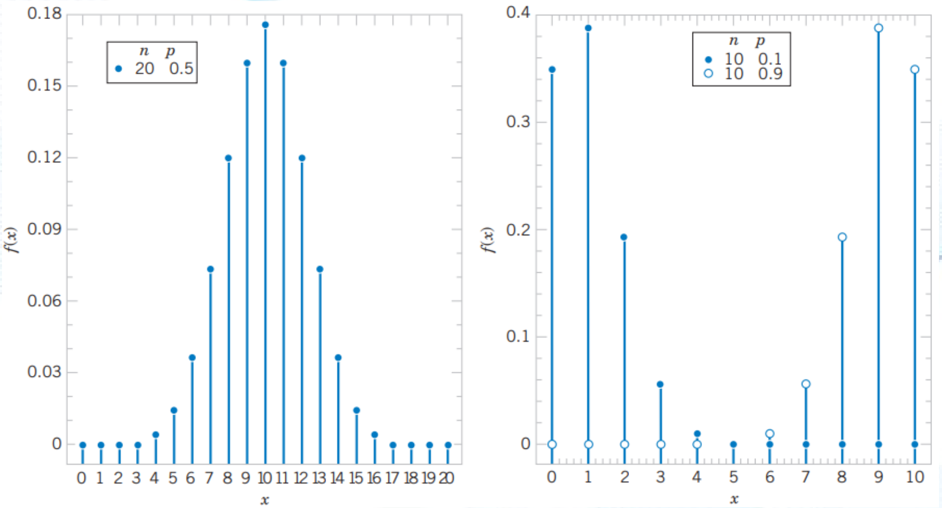


Figure. Binomial distributions for selected values of n and p.

Example

Flip a coin 5 times. Find the probability of the following cases:

1. Exactly 2 heads.
2. At least 4 heads.

Solution. The repeated tossing of the coin is an example of a Bernoulli trial.

- Number of trials: $n = 5$.
- Probability of head: $p = 0.5$ and hence the probability of tail, $1 - p = 0.5$.

1. For exactly 2 heads, i.e. $x = 2$ and so

$$P(X = 2) = \binom{5}{2} 0.5^2 0.5^3 = \frac{5}{16}.$$

2. For at least 4 heads, i.e. $x \geq 4$ and thus

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) \\ &= \binom{5}{4} 0.5^4 0.5 + \binom{5}{5} 0.5^5 0.5^0 \\ &= \frac{3}{16}. \end{aligned}$$

Quiz 1

A fair coin is tossed 10 times, what are the probabilities of getting exactly 6 heads and at least 6 heads.

Quiz 2

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant.Let X be the number of samples that contain the pollutant in the next 18 samples analyzed.

1. Find $P(X = 2)$.
2. Determine the probability that at least four samples contain the pollutant.
3. Determine the probability that $3 \leq X < 7$.
4. Find the mean and standard deviation of X.

Table of Contents

Geometric Distribution

In a series of Bernoulli trials (independent trials with constant probability p of a success), let a random variable X denote the number of trials until the first success occur. Then X is a **geometric random variable** with parameter $0 < p < 1$ and

$$f(x) = P(X = x) = (1 - p)^{x-1}p$$

where $x = 1, 2, \ldots$

Note

Cumulative distribution function is given by

$$P(X \leq x) = 1 - (1 - p)^{x}.$$

Mean and Variance

Let X be a geometric random variable with parameter p. Then,

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1-p}{p^2}.$$

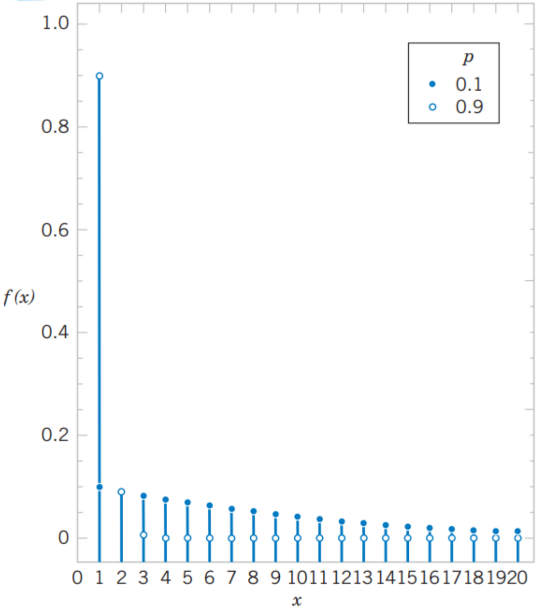


Figure. Geometric distributions for selected values of the parameter p.

$$P(X = x) = (1-p)^{x-1}p, \quad E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}.$$

Example

The probability that a wafer contains a large particle of contamination is 0.01. Assume that the wafers are independent, what is the probability that exactly 125 wafers need to be analyzed before a large particle is detected?

Solution. Let X denote the number of samples analyzed until a large particle is detected. Then, X is a geometric randomly variable with $p = 0.01$. The requested probability is

$$P(X = 125) = (1-0.01)^{125-1}0.01 = 0.0029.$$

Additional question. Determine the mean and variance of X.

Quiz

Suppose that the random variable X has a geometric distribution with $p = 0.5$. Determine the following probabilities:

1. $P(X = 4)$.
2. $P(X \leq 2)$.
3. $P(X > 2)$.

Compare the binomial and geometric distribution

Similarities:

1. The outcome of the experiments in both distributions can be classified as success or failure.
2. The probability of success is the same for each trial.
3. Each trial is independent.

Differences:

1. Binomial distribution:
   - There is a fixed number of trials.
     Example: Flip a coin 3 times.
   - For the random variable X, counts the number of successes in those trials.
2. Geometric distribution:
   - We are interested in the number of trials required until we obtain a success (i.e., there is no fixed number of trials).
     Example: How many flips will we need to make before we see Tails?
   - For the random variable X, counts the number of trials required to obtain that first success.

Select one answer for the each following quiz question:

1. Binomial distribution
2. Geometric distribution
3. Neither

Quiz 1

Jessica plays a game of luck in which she keeps rolling a dice until it lands on the number 4. Let X be the number of rolls until a 4 appears. What type of distribution does the random variable X follow?

Quiz 2

Henry makes 80% of all free-throws he attempts. Suppose he shoots 10 free-throws. Let X be the number of times Henry makes a basket during the 10 attempts. What type of distribution does the random variable X follow?

Quiz 3

A class with 30 students randomly selects a different student each week to bring a class snack. Of the students, 10% have food allergies. Let X be the number of weeks until a student with a food allergy is selected to bring the snack. What type of distribution does the random variable X follow?

Negative Binomial Distribution

In a series of Bernoulli trials (independent trials with constant probability p of a success), let a random variable X denote the number of trials until r success occur. Then X is a **negative binomial random variable** with parameters $0 < p < 1$ and $r = 1, 2, \ldots$, and

$$f(x) = P(X = x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r$$

where $x = r, r+1, r+2, \ldots$

Mean and Variance

Let X be a negative binomial random variable with parameter p and r. Then,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2}.$$

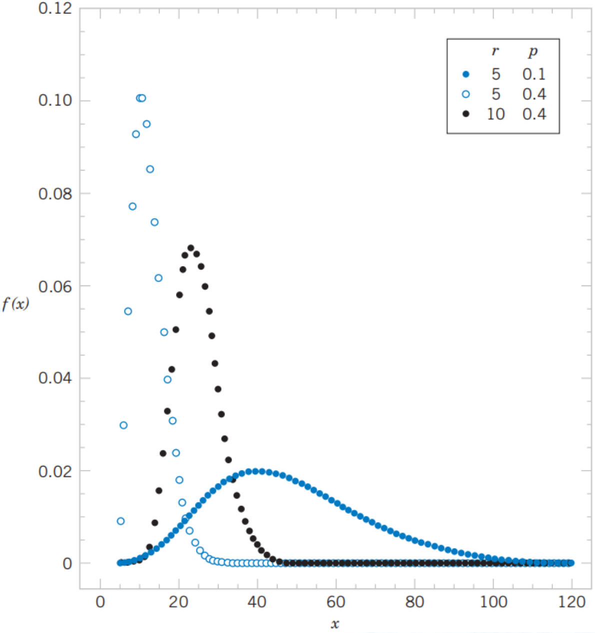


Figure. Negative binomial distributions for selected values of the parameters r and p.

FUQN

$$f(x) = P(X = x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r, \quad E(X) = \frac{r}{p}, \quad V(X) = \frac{r(1-p)}{p^2}.$$

Note

1. The geometric and negative binomial distributions are related to the binomial distribution in that the underlying probability experiment is the same, i.e., independent trials with two possible outcomes.
2. The random variable defined in the geometric and negative binomial case highlights a different aspect of the experiment, namely the number of trials needed to obtain a specific number of "successes".

Quiz

Suppose that X is a negative binomial random variable with $p = 0.2$ and $r = 4$. Determine the following:

1. $E(X)$.
2. $P(X = 20)$.

Table of Contents

Hyper-geometric Distribution

A set of N objects contains

- K objects classified as successes
- $N - K$ objects classified as failures.

A sample of size n objects is selected randomly (without replacement) from the N objects, where $K \leq N$ and $n \leq N$. Let a random variable X denote the number of successes in the sample. Then X is a **hyper-geometric random variable** and

$$f(x) = P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

where $x = \max\{0, n + K - N\}$ to $\min\{K, n\}$.

Note. We can write $X \sim H(N, K, n)$.

Mean and Variance

Let X be a hyper-geometric random variable with parameters N, K and n. Then,

$$\mu = E(X) = np \quad \text{and} \quad \sigma^2 = V(X) = np(1-p)\frac{N-n}{N-1}$$

where $p = \dfrac{K}{N}$.

Note

The term in the variance of a hyper-geometric random variable $\dfrac{N-n}{N-1}$ is called the **finite population correction factor**.

Example

Suppose your friend has 10 cookies, 3 of which are chocolate chip. Your friend randomly divides the cookies equally between herself and you. What is the probability that you get all the chocolate chip cookies?

Solution. Let X be the number of chocolate chip cookies you get. Then X follows hypergeometric distribution with

- $N = 10$ total cookies,
- $K = 3$ chocolate chip cookies,
- $n = 5$ cookies selected randomly by your friend to give to you.

We want the probability that you get all the chocolate chip cookies, i.e., $x = 3$ and

$$P(X = 3) = \frac{\binom{3}{3}\binom{7}{2}}{\binom{10}{5}} = 0.083.$$

Question: What is the probability that you get two the chocolate chip cookies?

Note that X has a hypergeometric distribution and not binomial because the cookies are being selected (or divided) without replacement.

$$f(x) = P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Quiz (Parts from Suppliers)

A batch of parts contains 100 parts from a local supplier of tubing and 200 parts from a supplier of tubing in the next state. If four parts are selected randomly and without replacement.

1. What is the probability they are all from the local supplier?
2. What is the probability that two or more parts in the sample are from the local supplier?
3. What is the probability that at least one part in the sample is from the local supplier?

Note

Sampling without replacement is frequently used for inspection and the hypergeometric distribution simplifies the calculations.

Table of Contents

Poisson Distribution

The random variable X that equals the number of events in a Poisson process is a **Poisson random variable** with parameter $0 < \lambda$, and the probability mass function of X is

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $x = 0, 1, 2, \ldots$

<u>**Note**</u>. We can write $X \sim$ Poisson(λ).

Mean and Variance

Let X be a Poisson random variable with parameter λ. Then,

$$\mu = E(X) = \lambda \quad \text{and} \quad \sigma^2 = V(X) = \lambda.$$

Note

1. The main application of the Poisson distribution is to count the number of times some event occurs over a fixed interval of time or space. For example,
   - The number of customers arriving at Big C between 9a.m. and 10a.m.
   - The number of calls made to 113 in Quy Nhon on a Saturday.
   - The number of accidents at a particular intersection during the month of April.
   - The number of typos occur per page in the Nhan Dan newspaper, which is 5 pages long.
2. The Poisson distribution is similar to all previously considered families of discrete probability distributions in that it counts the number of times something happens.
3. The Poisson distribution is different in that there is not an act that is being repeatedly performed.

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ where } x = 0, 1, 2, \ldots$$

Example (Calculations for Wire Flaws)

For the case of the thin copper wire, suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter. Determine the probability of

1. 10 flaws in 5 millimeters of wire.

   Solution. Let X denote the number of flaws in 5 millimeter of wire. Then, $E(X) = 5 \cdot 2.3 = 11.5$ flaws and

   $$P(X = 10) = \frac{e^{-11.5}11.5^{10}}{10!} = 0.113.$$

2. at least one flaw in 2 millimeters of wire.

   Solution. Let X denote the number of flaws in 2 millimeter of wire. Then, $E(X) = 2 \cdot 2.3 = 4.6$ flaws and

   $$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-4.6}4.6^0}{0!} = 0.9899.$$

Quiz (CDs)

Contamination is a problem in the manufacture of optical storage disks (CDs). The number of particles of contamination that occur on an optical disk has a Poisson distribution, and the average number of particles per centimeter squared of media surface is 0.1. The area of a disk under study is 100 squared centimeters. Find the probability that 12 particles occur in the area of a disk under study.

Quiz

Suppose that the number of customers who enter a bank in an hour is a Poisson random variable, and suppose that $P(X = 0) = 0.05$. Determine the mean and variance of X.

Thank you!