Advanced Probability and Statistics

MATH4010

February 18, 2025

Table of contents

Introduction of this course

Course Introduction: Advanced Probability & Statistics

- INSTRUCTOR: Chu Thi Mai Hong
- E-MAIL: hong.ctm@vinuni.edu.vn
- LEC: Wed 09:00 - 10:50 (C302)

- TA: Lieu Long Ho
- E-MAIL: ho.ll@vinuni.edu.vn
- REC: Fri 09:00 - 10:50 (C302)

Textbooks

- Probability, Statistics, and Random Processes for Electrical Engineering, Alberto Leon-Garcia
- MCMC from Scratch – A Practical Introduction to Markov Chain Monte Carlo, Masanori Hanada & So Matsuura

Course Instruction: Grading Scheme

The assignments will be graded on criteria as below.

Course Component	Overall Weight	Tentative Period
Attendance	5%	Week 1 - 15
Exams (In-class)		
Quiz 1	10%	Week 5
Midterm	25%	Week 9
Quiz 2	10%	Week 12
Project		
Report submission	25%	Week 14
Presentation	25%	Week 15

Using probability? Be careful!

Example

In the 1996 presidential race, Senator Bob Dole's age became an issue.

- Bob Dole is 72 years old.
- 72-year-old white male has a 27 percent risk of dying in the next four years.

Is there a problem?

Misuse of probability

Example

In the 1996 presidential race, Senator Bob Dole's age became an issue.

- Bob Dole is 72 years old.
- 72-year-old white male has a 27 percent risk of dying in the next four years.

Is there a problem?

The trouble with this argument is that the probability of survival was **not conditioned** on additional pertinent facts. As it happens, if a 72-year-old male is still in the workforce and, additionally, happens to be rich, then the average he has only a one-in-eight chance of dying in the next four years.

Misuse of probability

Example

In a sample of American women between the ages of 35 and 50, 4 out of 100 develop cancer within a year. Does Mrs. Smith, a 49-year-old American woman, there fore have a 4% chance of getting cancer in the next year?

Misuse of probability

VINUNIVERSITY

Example

In a sample of American women between the ages of 35 and 50, 4 out of 100 develop cancer within a year. Does Mrs. Smith, a 49-year-old American woman, there fore have a 4% chance of getting cancer in the next year?

There is no answer.

Example

In a sample of American women between the ages of 35 and 50, 4 out of 100 develop cancer within a year. Does Mrs. Smith, a 49-year-old American woman, there fore have a 4% chance of getting cancer in the next year?

There is no answer.

- Suppose that in a sample of women between the ages of 45 and 90, it is 11 out of 100 develop cancer in a year. Are Mrs. Smith's chances 4%, or are they 11%?

VINUNIVERSITY

Example

In a sample of American women between the ages of 35 and 50, 4 out of 100 develop cancer within a year. Does Mrs. Smith, a 49-year-old American woman, there fore have a 4% chance of getting cancer in the next year?

There is no answer.

- Suppose that in a sample of women between the ages of 45 and 90, it is 11 out of 100 develop cancer in a year. Are Mrs. Smith's chances 4%, or are they 11%?
- Suppose that her mother had cancer, and 22 out of 100 women between 45 and 90 whose mothers had the disease will develop it. Are her chances 4%, 11%, or 22%?

VINUNIVERSITY

Example

In a sample of American women between the ages of 35 and 50, 4 out of 100 develop cancer within a year. Does Mrs. Smith, a 49-year-old American woman, there fore have a 4% chance of getting cancer in the next year?

There is no answer.

- Suppose that in a sample of women between the ages of 45 and 90, it is 11 out of 100 develop cancer in a year. Are Mrs. Smith's chances 4%, or are they 11%?
- Suppose that her mother had cancer, and 22 out of 100 women between 45 and 90 whose mothers had the disease will develop it. Are her chances 4%, 11%, or 22%?
- You might think, the more specific the class, the better—but the more specific the class, the smaller its size and the less reliable the frequency. If there were only two people in the world very much like Mrs. Smith, and one developed cancer, would anyone say that Mrs. Smith's chances are 50%?

Generate discrete random variable

Generate discrete random variable

Most programming languages can deliver samples from the uniform distribution to us. The Python code for generating uniform random variables is:

```
rng = np.random.default_rng()
rng.random()
```

Example (Bernoulli)

Simulate tossing a coin with probability of heads p.

Generate discrete random variable

Most programming languages can deliver samples from the uniform distribution to us. The Python code for generating uniform random variables is:

```
rng = np.random.default_rng()
rng.random()
```

Example (Bernoulli)

Simulate tossing a coin with probability of heads p.

Solution: Let U be a Uniform$(0, 1)$ random variable. We can write a Bernoulli random variable X as:

$$X = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{if } U \geq p \end{cases}$$

Thus,

$$P(H) = P(X = 1) = P(U < p) = p.$$

Most programming languages can deliver samples from the uniform distribution to us. The Python code for generating uniform random variables is:

```
rng = np.random.default_rng()
rng.random()
```

Example (Bernoulli)

Simulate tossing a coin with probability of heads p.

Solution: Let U be a Uniform$(0, 1)$ random variable. We can write a Bernoulli random variable X as:

$$X = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{if } U \geq p \end{cases}$$

Thus,

$$P(H) = P(X = 1) = P(U < p) = p.$$

Example (Binomial)

Generate a Binomial(50,0.2) random variable.

Example

Give an algorithm to simulate the value of a random variable X such that

$$P(X = 2) = 0.15, \quad P(X = 3) = 0.4, \quad P(X = 4) = 0.1.$$

Solution: We divide the interval [0, 1] into subintervals as follows:

$$\begin{aligned} A_1 &= [0, 0.15) \\ A_2 &= [0.15, 0.55) \\ A_3 &= [0.55, 0.65) \\ A_4 &= [0.65, 1] \end{aligned}$$

Subinterval A_i has length p_i. We obtain a uniform number U. If U belongs to A_i, then $X = x_i$.

$$P(X = x_i) = P(U \in A_i) = p_i$$

Example

Generate a Poisson random variable with parameter $\lambda = 2$.

Solution: We know a Poisson random variable has

$$p_i = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!} \quad \text{for } i = 0, 1, 2, \ldots$$

Example

Generate a Poisson random variable with parameter $\lambda = 2$.

Solution: We know a Poisson random variable has

$$p_i = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!} \quad \text{for } i = 0, 1, 2, \dots$$

Next, we divide the interval $[0, 1)$ into subintervals such that the subinterval has length p_j.

$$X = \begin{cases} x_0 & \text{if } (U < p_0) \\ x_j & \text{if } \left(\sum_{k=0}^{j-1} p_k \leq U < \sum_{k=0}^{j} p_k\right) \end{cases}$$

Example

Generate a Poisson random variable with parameter $\lambda = 2$.

Solution: We know a Poisson random variable has

$$p_i = P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!} \quad \text{for } i = 0, 1, 2, \ldots$$

Next, we divide the interval $[0, 1)$ into subintervals such that the subinterval has length p_j.

$$X = \begin{cases} x_0 & \text{if } (U < p_0) \\ x_j & \text{if } \left(\sum_{k=0}^{j-1} p_k \leq U < \sum_{k=0}^{j} p_k\right) \end{cases}$$

Here $x_i = i - 1$, so

$$X = i \quad \text{if } p_0 + p_1 + \ldots + p_{i-1} < U < p_0 + p_1 + \ldots + p_i$$

$$F(i-1) = P(X \leq i-1) \text{ is c.d.f.}$$

Generate continuous random variable

Generate continuous random variable

Theorem

Let $U \sim \text{Uniform}(0, 1)$ and F be a c.d.f. which is strictly increasing. Also, consider a random variable X defined as

$$X = F^{-1}(U).$$

Then the c.d.f. of X is F.

$$X \sim F$$

Theorem

Let $U \sim \text{Uniform}(0,1)$ and F be a c.d.f. which is strictly increasing. Also, consider a random variable X defined as

$$X = F^{-1}(U).$$

Then the c.d.f. of X is F.

$$X \sim F$$

Proof:

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

Example

Generate an Exponential(1) random variable.

Example

Generate an Exponential(1) random variable.

Solution: To generate an Exponential random variable with parameter $\lambda = 1$, we proceed as follows:

$$F(x) = 1 - e^{-x} \text{ for } x > 0 \text{ and } 0 \text{ otherwise.}$$

$$U \sim \text{Uniform}(0, 1)$$

$$X = F^{-1}(U) = -\ln(1 - U)$$

Example

Generate an Exponential(1) random variable.

Solution: To generate an Exponential random variable with parameter $\lambda = 1$, we proceed as follows:

$$F(x) = 1 - e^{-x} \text{ for } x > 0 \text{ and } 0 \text{ otherwise.}$$

$$U \sim \text{Uniform}(0, 1)$$

$$X = F^{-1}(U) = -\ln(1 - U)$$

This formula can be simplified since

$$(1 - U) \sim \text{Uniform}(0, 1)$$

Thus,

$$X = -\ln(U)$$

Example

Generate a random variable having c.d.f.

$$F(x) = \frac{x^2 + x}{2}, \quad 0 \leq x \leq 1.$$

Example

Generate a random variable having c.d.f.

$$F(x) = \frac{x^2 + x}{2}, \quad 0 \leq x \leq 1.$$

Solution:

Set

$$X^2 + X = 2U,$$

where U is a standard uniform random variable.

Example

Generate a random variable having c.d.f.

$$F(x) = \frac{x^2 + x}{2}, \quad 0 \le x \le 1.$$

Solution:

Set

$$X^2 + X = 2U,$$

where U is a standard uniform random variable.

Then,

$$(X + \frac{1}{2})^2 = 2U + \frac{1}{4} \implies X + \frac{1}{2} = \sqrt{2U + \frac{1}{4}}.$$

Thus,

$$X = \sqrt{2U + \frac{1}{4}} - \frac{1}{2} \quad (X, U \in [0, 1]).$$

Example

Generate a random variable with the density

$$f(x) = 2.5x\sqrt{x} \quad \text{for } 0 < x < 1$$

Example

Generate a random variable with the density

$$f(x) = 2.5x\sqrt{x} \quad \text{for } 0 < x < 1$$

Solution:

$$F_X(a) = \int_0^a f(x)\,dx = \int_0^a 2.5x\sqrt{x}\,dx = \int_0^a 2.5x^{3/2}\,dx = \frac{5}{8}a^{\frac{5}{2}}.$$

Example

Generate a random variable with the density

$$f(x) = 2.5x\sqrt{x} \quad \text{for } 0 < x < 1$$

Solution:

$$F_X(a) = \int_0^a f(x)\,dx = \int_0^a 2.5x\sqrt{x}\,dx = \int_0^a 2.5x^{3/2}\,dx = \frac{5}{8}a^{\frac{5}{2}}.$$

Thus,

$$\frac{5}{8}X^{\frac{5}{2}} = U \implies X^{\frac{5}{2}} = \frac{8}{5}U \implies X = \left(\frac{8}{5}U\right)^{\frac{2}{5}}.$$

Generate continuous random variable

When $F(x)$ doesn't have a simple form but the p.d.f. $f(x)$ is available, random variables with density $f(x)$ can be generated by the rejection method.

Suppose you have a method for generating a random variable having density function $g(y)$. Now, assume you want to generate a random variable having density function $f(x)$. Let c be a constant such that

$$\frac{f(y)}{g(y)} \leq c \quad \text{(for all } y\text{)}.$$

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \leq c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \leq \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \leq c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \leq \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

VINUNIVERSITY

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \leq c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \leq \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

Sketch of Solution: The number of times N that the first two steps of the algorithm need to be called is a geometric r.v. with $p = P\left(U \leq \frac{f(Y)}{cg(Y)}\right)$.

Generate continuous random variable

VINUNIVERSITY

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \le c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \le \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

Sketch of Solution: The number of times N that the first two steps of the algorithm need to be called is a geometric r.v. with $p = P\left(U \le \frac{f(Y)}{cg(Y)}\right)$.

Thus, $E[N] = P\left(U \le \frac{f(Y)}{cg(Y)}\right) = \int_0^1 g(y) \cdot P\left(U \le \frac{f(y)}{cg(y)} \mid Y = y\right) dy$.

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \leq c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \leq \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

Sketch of Solution: The number of times N that the first two steps of the algorithm need to be called is a geometric r.v. with $p = P\left(U \leq \frac{f(Y)}{cg(Y)}\right)$.
Thus, $E[N] = P\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_0^1 g(y) \cdot P\left(U \leq \frac{f(y)}{cg(y)} \mid Y = y\right) dy$. Let $M = \{U \leq \frac{f(Y)}{cg(Y)}\}$.
Then $P(Y \in A \mid M) = \frac{P(Y \in A \cap M)}{P(M)}$.

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \leq c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \leq \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

Sketch of Solution: The number of times N that the first two steps of the algorithm need to be called is a geometric r.v. with $p = P\left(U \leq \frac{f(Y)}{cg(Y)}\right)$.

Thus, $E[N] = P\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_0^1 g(y) \cdot P\left(U \leq \frac{f(y)}{cg(y)} \mid Y = y\right) dy$. Let $M = \{U \leq \frac{f(Y)}{cg(Y)}\}$.

Then $P(Y \in A \mid M) = \frac{P(Y \in A \cap M)}{P(M)}$. Besides, $P\left(Y \leq y \mid M\right) = \frac{P(Y \leq y \cap M)}{P(M)}$.

VINUNIVERSITY

Rejection method

Input target density f, available Y with density g, and $\frac{f(y)}{g(y)} \le c$ (for all y).

- Generate a random variable Y according to the density function $g(y)$.
- Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
- Accept Y if $U \le \frac{f(Y)}{cg(Y)}$; otherwise, reject Y and repeat the process.

Output X such that its density is f.

Show that rejection method is valid.

Sketch of Solution: The number of times N that the first two steps of the algorithm need to be called is a geometric r.v. with $p = P\left(U \le \frac{f(Y)}{cg(Y)}\right)$.

Thus, $E[N] = P\left(U \le \frac{f(Y)}{cg(Y)}\right) = \int_0^1 g(y) \cdot P\left(U \le \frac{f(y)}{cg(y)} \mid Y = y\right) dy$. Let $M = \{U \le \frac{f(Y)}{cg(Y)}\}$.

Then $P(Y \in A \mid M) = \frac{P(Y \in A \cap M)}{P(M)}$. Besides, $P\left(Y \le y \mid M\right) = \frac{P(Y \le y \cap M)}{P(M)}$.

Thus, we get $P\left(Y \le y \mid M\right) = \frac{P\left(U \le \frac{f(Y)}{cg(Y)}\right)}{P(M)}$.

By the law of total probability, $P(K \mid M) = P(K) \cdot \frac{P(M|K)}{P(M)}$.

Thus: $P(K \mid M) = \frac{P(Y \le y)}{P(M)} = \frac{F(y)}{P(M)}$.