Семплирование в малых размерностях.

Елена Юрьевна Шмилева

2 ноября 2018 г.

Литература.

- Sheldon Ross, 2013, "Simulation" 5th Edition, Academic Press, 2013
- М.Б. Лагутин, "Наглядная статистика", 2018, бе издание
- Steven S. Skiena, "Data Science Design", 2017, Springer
- J.E.Gentle, 2009, Computational Statistics, Springer, 720 pages
- Г.И.Ивченко, Ю.И.Медведев, Введение в математическую статистику, Издательство ЛКИ, Москва, 2014, 600 с.

🕕 Случайные величины

Генераторы случайных чисел

Семплирование. Метод обратного преобразования

4 Семплирование. Метод выборки с отклонением 🚁 🔻 🔻 🗦 🔻 🗨 🗨

План

- Случайные величины
- 2 Генераторы случайных чисел
- ③ Семплирование. Метод обратного преобразования
- 4 Семплирование. Метод выборки с отклонением

Случайная величина

Пусть проводится некий случайный эксперимент на вероятностном пространстве (Ω, \mathcal{F}, P) .

Что такое случайная величина?



Случайная величина

Пусть проводится некий случайный эксперимент на вероятностном пространстве (Ω, \mathcal{F}, P) .

Что такое случайная величина?

Определение

Функция $X:\Omega \to {f R}$ называется случайной величиной, если для любых вещественных чисел а и b множество исходов ω , таких что $X(\omega) \in (a,b)$, принадлежит ${\cal F}.$

Типы распределений случайных величин.

- Дискретные случайные величины.
- Абсолютно непрерывные случайные величины.
- Дискретно-непрерывные (смешанные).

Дискретная случайная величина

Определение

Случайная величина X называется дискретной, если она принимает конечное или счетное множество значений.

- •
- •

- 0

- Бернулли.
- •
- •
- •

- Бернулли.
- Биномиальное.
- •
- •
- 0
- 0

- Бернулли.
- Биномиальное.
- Гипергеометрическое.
- •

- Бернулли.
- Биномиальное.
- Гипергеометрическое.
- Пуассоновское.
- 0

- Бернулли.
- Биномиальное.
- Гипергеометрическое.
- Пуассоновское.
- Геометрическое.
- 0

- Бернулли.
- Биномиальное.
- Гипергеометрическое.
- Пуассоновское.
- Геометрическое.
- и т.д.

Непрерывная случайная величина.

Определение

Случайная величина X называется (абсолютно) непрерывной, если X принимает все значения из некоторого интервала.

- •
- •

- Равномерная случайная величина
- •
- 0
- •
- •

- Равномерная случайная величина
- Экспоненциальная случайная величина
- •
- •
- •

- Равномерная случайная величина
- Экспоненциальная случайная величина
- Нормальная (гауссовская) случайная величина
- 0

- Равномерная случайная величина
- Экспоненциальная случайная величина
- Нормальная (гауссовская) случайная величина
- Бета случайная величина

0

- Равномерная случайная величина
- Экспоненциальная случайная величина
- Нормальная (гауссовская) случайная величина
- Бета случайная величина
- Коши, Парето и т.д.

Задание закона случайной величины

Какой есть универсальный способ задания распределения любой случайной величины (случайного вектора)?

Задание закона случайной величины

Какой есть универсальный способ задания распределения любой случайной величины (случайного вектора)?

через функцию распределения

Функция распределения.

Определение

Функцией распределения (cumulative distribution function) случайной величины X называется следующая функция: $F(a) = P(X \le a)$.

Функция распределения есть и у дискетных, и у непрерывных случайных величин.

Замечание:

- •
- •
- a

Замечание:

• непрерывная справа функция

$$F(a+0) = \lim_{x \to a+0} F(x) = F(a)$$

- •
- 0

Замечание:

• непрерывная справа функция

$$F(a+0) = \lim_{x \to a+0} F(x) = F(a)$$

- для дискретных случайных величин ступенчатая cádlág функция,
- •

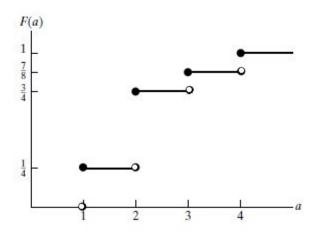
Замечание:

• непрерывная справа функция

$$F(a+0) = \lim_{x \to a+0} F(x) = F(a)$$

- для дискретных случайных величин ступенчатая cádlág функция,
- для непрерывных случайных величин непрерывная неубывающая функция.

Функция распределения. Пример для дискретных случайных величин.



Симуляции

Определение

Симуляции (или семплинг) - это способ определять вероятности посредством компьютерного эксперимента.

Симуляции сл.величин

С чего начать?



Симуляции сл. величин

С чего начать?

• Генерация последовательности равномерно распределенных случайных величин.

Симуляции сл. величин

С чего начать?

- Генерация последовательности равномерно распределенных случайных величин.
- Генерация последовательности любых других случайных величин/векторов.

Симуляции сл. величин

С чего начать?

- Генерация последовательности равномерно распределенных случайных величин.
- Генерация последовательности любых других случайных величин/векторов.
- Симуляция событий, оценка вероятностей и т.д.

План

- Случайные величины
- Генераторы случайных чисел
- ③ Семплирование. Метод обратного преобразования
- 4 Семплирование. Метод выборки с отклонением

Генераторы случайных чисел

Определение

Код, производящий последовательность псевдослучайных чисел, называется генератором случайных чисел.

Псевдослучайные числа

Определение

Псевдослучайные числа - это числа похожие по своим свойствам на равномерно распределенные случайные на отрезке [0,1], но полученные детерминистическим способом.

- 0
- •
- •

- Метод середины квадрата.
- •
- •

- Метод середины квадрата.
- Много других генераторов.
- •

Генераторов случайных чисел много:

- Метод середины квадрата.
- Много других генераторов.
- МТ19937 используется в R по умолчанию. Mersenne twister МТ19937 период простое число Мерсенна $2^{19937}-1$.

•

- Метод середины квадрата.
- Много других генераторов.
- МТ19937 используется в R по умолчанию. Mersenne twister МТ19937 период простое число Мерсенна $2^{19937} 1$.
- Линейный Конгруентный Генератор (мультипликативный датчик) прост для понимания.

Линейный Конгруентный Генератор (LCG)

Самый простой генератор случайных чисел - это Линейный Конгруентный Генератор (LCG) или мультипликативный датчик

Линейный Конгруентный Генератор (LCG)

Самый простой генератор случайных чисел - это Линейный Конгруентный Генератор (LCG) или мультипликативный датчик

$$X_n = (aX_{n-1} + c) \mod m,$$

$$U_n = X_n/m,$$

где *a*, *c*, *m* - целые числа

- a > 0;
- $c \ge 0$;
- $m \geq 0$;
- X_1 seed (зерно) изначально заданное целое число;
- $X_1 < m$.

Рекурентная формула:

$$X_n = (7X_{n-1} + 7) \mod 10,$$

$$X_1 = 7$$

Рекурентная формула:

$$X_n = (7X_{n-1} + 7) \mod 10,$$

 $X_1 = 7$

Получим Х: 7, 6, 9, 0, 7, 6, 9, 0,...

Получим U: 0.7, 0.6, 0.9, 0, 0.7, 0.6, 0.9, 0,...

Период = 4.

Рекурентная формула:

$$X_n = (6X_{n-1} + 7) \mod 23,$$

$$X_1 = 5$$

Рекурентная формула:

$$X_n = (6X_{n-1} + 7) \mod 23,$$

 $X_1 = 5$

Получаем Х: 5, 14, 22, 1, 13, 16, 11, 4, 9, 15, 5 ...

Получаем U: 5/23, 14/23, 22/23, 1/23, 13/23, 16/23, 11/23, 4/23, 9/23, 15/23, 5/23

15/23, 5/23 ...

Период = 11.

Замечания

Замечания:

- Как только $X_n = seed$ все зациклится;
- Длина периода не превышает m;
- т надо брать большим;
- "хороший" LCG помогает выбрать теория чисел.
- Генерирует лишь дробно-рациональные числа. Иррациональные числа получаются аппроксимацией.

Длина периода. Правило выбора "хорошего" LCG

Teopeмa (Hull& Dobell'1962, Knuth)

Длина периода равна т тогда и только тогда, когда:

- Числа с и т взаимно простые;
- b=a-1 кратно р для каждого простого р, являющегося делителем m;
- b кратно 4, если т кратно 4.
- \bullet Например, a=16, c=7, m=225 даст период 225
- Из теоремы видно, что хорошо брать m большим простым числом, с взаимно простым m, при этом a=1. Например, простое число Мерсена $m=2^{31}-1$.

Замечания

- Полный период еще не гарантирует, что генератор "хороший". Его нужно еще проверить на случайность всевозможными тестами.
- На практике действует "презумпция случайности": алгоритм используют, если не установлено, что он "плохой".

Вопрос

Как генерируют все остальные случайные величины и случайные векторы?

Вопрос

Как генерируют все остальные случайные величины и случайные векторы?

Различные вероятностные методы существуют для этого.

План

- Случайные величины
- 2 Генераторы случайных чисел
- Оемплирование. Метод обратного преобразования
- 4 Семплирование. Метод выборки с отклонением

Метод обратного преобразования

Пусть X с.в. с функцией распределения F

Fact

Если функция распределения F непрерывная и строго возрастающая $(\tau.e. \; \exists \; F^{-1} \;)$, то

$$\xi = F^{-1}(\eta),$$

где η равномерная с.в. на [0,1], определена и имеет такое же распределение как X.

Метод обратного преобразования (Continious Inverse Transform)

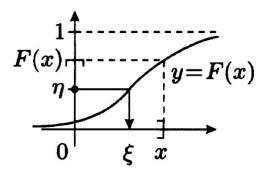


Рис.: График взят из книги Лагутин, Наглядная Статистика

η имитирует вероятнсть

ξ имитирует η-квантиль распределения F

Метод обратного преобразования (Continuous Inverse Transform)

Доказательство.

Найдем функцию распределения ξ

$$P\{\xi \le t\} = P\{F^{-1}(\eta) \le t\} =$$

$$= P\{\eta \le F(t)\} = F(t),$$

последнее равенство верно т.к. η равномерная на [0,1]. Значит, $Law(\xi) = F = Law(X)$.



Пример. Генерирование экспоненциального распределения

$$X\sim Exp(\lambda)$$
 с функцией распределения $F(x)=1-e^{-\lambda x}$, $x>0$. Найдем $F^{-1}(u)=-rac{1}{\lambda}\log(1-u)$

- Генерируем последовательность равномерных с.в. $U_1, U_2, ...;$
- Положим $X_n = -\frac{1}{\lambda} \log U_n$;
- $X_1, X_2, ... X_n, ...$ это последовательность экспоненциальных с.в. с параметром λ .

Пусть X дискретная с.в. с функцией распределения F Определим обобщенную обратную функцию

$$F^{-1}(y) = min\{x : F(x) \ge y\}, y \in [0, 1].$$

Fact

Пусть U равномерная с.в. на [0,1], тогда

$$Y = F^{-1}(U)$$

определена и имеет такое же распределение как X.

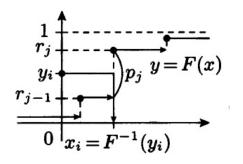


Рис.: График взят из книги Лагутин, Наглядная Статистика

Здесь $\{r_j\}$ накопленные вероятности, т.е. $r_j = p_1 + p_2 + \dots + p_j$, $r_0 = 0$

 $X \sim Pois(3)$

- ullet Генерируем последовательность равномерных с.в. $U_1,\,U_2,...$;
- $U_i \in (0; e^{-3}]$, to $X_i = F^{-1}(U_i) = 0$

$$X \sim Pois(3)$$

- Генерируем последовательность равномерных с.в. $U_1, U_2, ...;$
- $U_i \in (0; e^{-3}]$, to $X_i = F^{-1}(U_i) = 0$ $U_i \in (e^{-3}; e^{-3} + 3 \cdot e^{-3}]$, to $X_i = F^{-1}(U_i) = 1$

 $X \sim Pois(3)$

- Генерируем последовательность равномерных с.в. $U_1, U_2, ...;$
- $U_i \in (0; e^{-3}]$, to $X_i = F^{-1}(U_i) = 0$
- $U_i \in (e^{-3}; e^{-3} + 3 \cdot e^{-3}]$, to $X_i = F^{-1}(U_i) = 1$
- $U_i \in (e^{-3} + 3 \cdot e^{-3}; \ e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3}]$, to $X_i = F^{-1}(U_i) = 2$

 $X \sim Pois(3)$

- ullet Генерируем последовательность равномерных с.в. $U_1,\,U_2,...;$
- $U_i \in (0; e^{-3}]$, to $X_i = F^{-1}(U_i) = 0$
- $U_i \in (e^{-3}; e^{-3} + 3 \cdot e^{-3}]$, to $X_i = F^{-1}(U_i) = 1$
- $U_i \in (e^{-3} + 3 \cdot e^{-3}; e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3}]$, to $X_i = F^{-1}(U_i) = 2$
- $U_i \in (e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3}; e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3} + \frac{3^3}{3!}e^{-3}]$, to $X_i = F^{-1}(U_i) = 3$

 $X \sim Pois(3)$

- ullet Генерируем последовательность равномерных с.в. $U_1,\,U_2,...$;
- $U_i \in (0; e^{-3}]$, to $X_i = F^{-1}(U_i) = 0$
- $U_i \in (e^{-3}; e^{-3} + 3 \cdot e^{-3}]$, to $X_i = F^{-1}(U_i) = 1$
- $U_i \in (e^{-3} + 3 \cdot e^{-3}; e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3}]$, to $X_i = F^{-1}(U_i) = 2$
- $U_i \in (e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3}; e^{-3} + 3 \cdot e^{-3} + \frac{3^2}{2}e^{-3} + \frac{3^3}{3!}e^{-3}]$, то $X_i = F^{-1}(U_i) = 3$
- etc.



Моделирование распределения Бернулли.

 $X_1, X_2, ... X_n, ...$ последовательность независимых Бернуллиевских с.в.

$$P{X_n = 0} = p, P{X_n = 1} = 1 - p, p \in (0, 1).$$

• Генерируем последовательность равномерно распределенных с.в. $U_1, U_2, ..., U_n, ...$

Моделирование распределения Бернулли.

 $X_1, X_2, ..., X_n,$ последовательность независимых Бернуллиевских с.в.

$$P{X_n = 0} = p, P{X_n = 1} = 1 - p, p \in (0, 1).$$

- Генерируем последовательность равномерно распределенных с.в. $U_1, U_2, ..., U_n, ...$
- Положим

$$X_n = \begin{cases} 0, & U_n \leq p, \\ 1, & U_n > p. \end{cases}$$

Моделирование распределения Бернулли.

 $X_1, X_2, ... X_n, ...$ последовательность независимых Бернуллиевских с.в.

$$P{X_n = 0} = p, P{X_n = 1} = 1 - p, p \in (0, 1).$$

- Генерируем последовательность равномерно распределенных с.в. $U_1, U_2, ..., U_n, ...$
- Положим

$$X_n = \begin{cases} 0, & U_n \le p, \\ 1, & U_n > p. \end{cases}$$

• $P\{X_n = 0\} = P\{U_n < p\} = p$ последнее равенство из формулы для функции распределения равномерной случайной величины.

План

- Случайные величины
- 2 Генераторы случайных чисел
- ③ Семплирование. Метод обратного преобразования
- Семплирование. Метод выборки с отклонением

Недостатки предыдущих методов

- Обратную функцию не всегда можно явно найти (например, для нормального распределения)
- Для большинства случайных векторов сложно находить обратную функцию распределения.

Метод выборки с отклонением

Генерируем $X \sim F$, для которого F^{-1} явно не решается.

Ищем $Y \sim G$ такую, что

• Ү легко генерируется

0

$$\exists c > 0: f(t) \le c \cdot g(t) \tag{1}$$

для $t\in (-\infty,+\infty)$, где $f(\cdot)=F'(\cdot)$ и $g(\cdot)=G'(\cdot)$ плотности.

Алгоритм выборки с отклонением

Алгоритм выборки с отклонением (Accept/Reject method):

- *U_i* ~ *Unif* (0, 1)
- $Y_i \sim G$
- Если $U_i \leq \frac{f(Y_i)}{c \cdot g(Y_i)}$, то $X_i = Y_i$, иначе шаг 1.

Метод выборки с отклонением. Обоснование.

Fact

Если g плотность Y, f плотность F и c из условия (1), U и Y независимы, то

$$F(t) = P\left(Y \le t \mid U \le \frac{f(Y)}{c \cdot g(Y)}\right).$$

Какова эффективность алгоритма? Какой процент реализаций отклоняется?

Какова эффективность алгоритма? Какой процент реализаций отклоняется?

ullet P(успех цикла) $=P(U\leq rac{f(Y)}{cg(Y)})=\int_{-\infty}^{\infty}\int_{0}^{rac{f(t)}{cg(t)}}ds\;g(t)dt=rac{1}{c}$

Какова эффективность алгоритма? Какой процент реализаций отклоняется?

- ullet P(успех цикла) $=P(U\leq rac{f(Y)}{cg(Y)})=\int_{-\infty}^{\infty}\int_{0}^{rac{f(t)}{cg(t)}}ds\;g(t)dt=rac{1}{c}$
- Пусть ξ количество неудач до первого успеха, i.e. $\xi \sim Geom(p)$, p=1/c. Известно, что $E\xi=c$.

Какова эффективность алгоритма? Какой процент реализаций отклоняется?

- ullet P(успех цикла) $=P(U\leq rac{f(Y)}{cg(Y)})=\int_{-\infty}^{\infty}\int_{0}^{rac{f(t)}{cg(t)}}ds\;g(t)dt=rac{1}{c}$
- Пусть ξ количество неудач до первого успеха, i.e. $\xi \sim Geom(p)$, p=1/c. Известно, что $E\xi=c$.
- Т.е. в среднем происходит с неудач на 1 успех.

Сгенерировать iid из $Z \sim N(0,1)$, где плотность

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbf{R}.$$

• Начнем с X := |Z|

$$f(x) = \frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad x \in (0, +\infty).$$

• Будем применять метод выборки с отклонением.

Сгенерировать iid из $Z \sim N(0,1)$, где плотность

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbf{R}.$$

• Начнем с X := |Z|

$$f(x) = \frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad x \in (0, +\infty).$$

• Будем применять метод выборки с отклонением.

• $Y \sim Exp(1)$ с плотностью $g(x) = e^{-x}, x > 0$ (легко генерируется с помощью Непрерывного Обратного Преобразования).

•

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2e}{\pi}} \exp\{-(x-1)^2/2\} \le \sqrt{\frac{2e}{\pi}}$$

ullet Значит, $c=\sqrt{rac{2e}{\pi}}pprox 1.32$

$$\frac{f(x)}{cg(x)} = \exp\{-(x-1)^2/2\}$$

Генерация модуля нормальной с.в.

Далее алгоритм выборки с отклонением (Accept/Reject method):

- *U_i* ∼ *Unif* (0, 1)
- $Y_i \sim Exp(1)$
- Если $U_i \leq \exp\{-(Y_i-1)^2/2\}$, то $X_i = Y_i$, иначе шаг 1.

Умея генерировать |Z|, как сгенерировать $Z \sim N(0,1)$?

Умея генерировать |Z|, как сгенерировать $Z \sim N(0,1)$?

- ullet Генерируем $U \sim Unif(0,1)$
- 0

$$Z_i = \begin{cases} X_i & U \le 1/2, \\ -X_i, & U > 1/2. \end{cases}$$

Замечания

- Вероятность принятия: $\frac{1}{C}$ (т.е. 1 из C=1,32 реализаций будут приняты).
- $T \sim N(a, \sigma^2)$ получается преобразованием $T = \sigma Z + a$.
- Вспомогательное распределение должно
 - иметь такой же носитель,
 - хвосты того же порядка,
 - хорошо повторять пики (иначе моды не будут заметны),
 - константа С должна быть минимально возможной.

Замечания еще.

- В гауссовском случае хорошо работает для генерирования случайных векторов.
- Однако это сложная задача для векторов большой размерности (процент отклонения растет экспоненциально с увеличением размерности).

Гауссовские векторы

Пусть случайный вектор $\xi=(\xi_1,\dots,\xi_d)$ имеет вектор средних $\mathbf{a}=\mathsf{E}\xi$ и невырожденную матрицу ковариаций Σ , составленную из элементов $\Sigma_{ij}=\mathit{Cov}(\xi_i,\xi_j).$

Определение

Говорят, что вектор ξ имеет нормальное распределение $\mathcal{N}(a,\Sigma)$ в \mathbf{R}^d , если плотность этого вектора равна

$$f_{\xi}(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|det\Sigma|}} \exp\{-(1/2)(x-a)^T \Sigma^{-1}(x-a)\},$$

где $x \in \mathbf{R}^d$.

- 4 ロ ト 4 個 ト 4 差 ト 4 差 ト - 差 - 约 Q (^)

Гауссовский вектор с независимыми компонентами.

Компоненты вектора независимы, если $\Sigma = E$, где E - это единичная матрица.

Вектор ξ имеет стандартное нормальное распределение $\mathcal{N}(0,E)$ в \mathbf{R}^d , если плотность этого вектора равна

$$f_{\xi}(x_1,...x_d) = \frac{1}{\sqrt{2\pi}} \exp\{-x_1^2/2\}...\frac{1}{\sqrt{2\pi}} \exp\{-x_d^2/2\}.$$

Моделирование гауссовского вектора с независимыми компонентами

• Генерировать $|\xi| = (|\xi_1|, |\xi_2|, ..., |\xi_d|)$

$$f_{|\xi|}(x_1,...x_d) = \frac{2}{\sqrt{2\pi}} \exp\{-x_1^2/2\}...\frac{2}{\sqrt{2\pi}} \exp\{-x_d^2/2\}, \ x_i > 0.$$

• Многомерное экспоненциальное с независимыми компонентами

$$g(x_1,...x_d) = \exp\{-x_1\}...\exp\{-x_d\}, x_i > 0$$

легко генерируется покомпонентно.

$$c_d = \max_{x \in (\mathbf{R}^+)^d} \frac{f(x)}{g(x)} = \frac{2^{d/2}}{\pi^{d/2}} e^{d/2} = (1.32)^d \to_{d \to \infty} \infty$$

 Процент отклонения растет экспоненциально с увеличением размерности.

- •
- •
- •

- Линейные Конгруентные генераторы равномерных сл.в.
- •
- •
- •

- Линейные Конгруентные генераторы равномерных сл.в.
- Метод обратной функции для непрерывных сл.в.
- •
- •

Что мы сегодня изучили?

- Линейные Конгруентные генераторы равномерных сл.в.
- Метод обратной функции для непрерывных сл.в.
- Метод обратной функции для дискретных сл.в.

•

- Линейные Конгруентные генераторы равномерных сл.в.
- Метод обратной функции для непрерывных сл.в.
- Метод обратной функции для дискретных сл.в.
- Метод выборки с отклонением.

Спасибо за внимание!