

Лекция 14. Процесс Бернулли

12 мая 2022 г.

1 Процесс Бернулли

После сложных мартингалов мы рассмотрим более простой процесс Бернулли (для подготовки к рассмотрению более сложного процесса Пуассона). Процессом Бернулли называется случайный процесс $\{X_t\}_{t \in \mathbb{N}}$, в котором все X_t независимы и следуют одинаковому распределению Бернулли с одинаковым параметром p . С его помощью часто моделируют разные реальные штуки:

- Браки на производстве
- Прибытие посетителя в магазин в интервалы времен
- Поступление запросов на сервер

Как вообще стоит рассматривать случайный процесс, в том числе процесс Бернулли? В случае, когда у нас конечное множество с.в. $\{X_t\}$, то нам достаточно задать совместную функцию вероятности или совместную плотность вероятности для всех возможных подмножеств $\{X_{t_1}, \dots, X_{t_k}\}$. Это необходимо, так как с.в. могут быть и зависимы. Но в случае с бесконечными последовательностями даже если мы определим совместные функции/плотности вероятности на всех конечных подмножествах, мы все равно не сможем определить, например, вероятность того что все $X_t = 1$.

Поэтому в случае со случайными процессами стоит рассматривать Ω как множество всех возможных последовательностей, которые могут получиться из значений случайных величин. В случае с процессом Бернулли — это все возможные последовательности из нулей и единиц. В таком случае легко найти вероятность события, что все $X_t = 1$ как в случае, когда $p < 1$:

$$\begin{aligned}\Pr(\forall t \in \mathbb{N} X_t = 1) &\leq \Pr(\forall t \in [1..n] X_t = 1) = p^n \\ &\Rightarrow \Pr(\forall t \in \mathbb{N} X_t = 1) = 0,\end{aligned}$$

так и для $p = 1$:

$$\Pr(\forall t \in \mathbb{N} X_t = 1) \geq 1 - \sum_{t=1}^{+\infty} \Pr[X_t = 0] = 1.$$

В дальнейшем случай с $p = 1$ (как и с $p = 0$) мы рассматривать не будем, так как ничего интересного в этом нет.

1.1 Свойства процесса Бернулли

Рассмотрим несколько свойств процесса Бернулли. Напомним, что так как каждый X_t следует распределению Бернулли независимо от других с.в., то

- $E[X_t] = p$
- $\text{Var}(X_t) = p(1 - p)$

Рассмотрим следующие две случайные величины.

S_n — **число единиц в первые n единиц времени**. Так как X_t независимы, то $S_n \sim \text{Bin}(n, p)$, причем

- $\Pr[S_n = k] = \binom{n}{k} p^k (1 - p)^{n-k}$
- $E[S_n] = np$
- $\text{Var}(S_n) = np(1 - p)$

T_1 — **время первой единицы**. Так как X_t независимы, то $T_1 \sim \text{Geom}(p)$, причем

- $\Pr[T_1 = k] = p(1 - p)^{k-1}$
- $E[T_1] = \frac{1}{p}$
- $\text{Var}(T_1) = \frac{1-p}{p^2}$

Беспамятство. Если мы рассматриваем процесс $Y_t = X_{t+n}$ для какого-то фиксированного n , то это опять будет процесс Бернулли, так как:

- Все Y_t независимы
- Все Y_t имеют распределение $\text{Bern}(p)$

Более того, если мы говорим, что $Y_t = X_{t+N}$, где N — какая-то случайная величина, то в некоторых случаях мы также получаем процесс Бернулли. Например:

- Пусть N — время первого успеха в процессе X_t . Тогда все X_{t+N} независимы друг от друга и от N , поэтому являются процессом Бернулли.
- Пусть теперь N — время перед первым успехом. Тогда X_{N+1} уже не независим от N , а точно равен единице.

В общем случае Y_t будет процессом Бернулли, если N является временем останова, то есть событие $N \geq n$ зависит только от X_1, \dots, X_{n-1} .

Беспамятство дает нам много других полезных свойств процесса Бернулли.

Продолжительность занятого периода. Рассмотрим сервер, на который в каждую единицу времени приходит или не приходит запрос. Это можно описать процессом Бернулли, где событие “во время t пришел запрос” соответствует $X_t = 1$. Скажем, что сервер занят во время t , если в это время на него пришел запрос. Занятый период — это отрезок $[t_1..t_2]$, такой, что во все $t \in [t_1..t_2]$ сервер занят, но в моменты $t_1 - 1$ и $t_2 + 1$ — свободен или еще не был запущен. Нас интересует продолжительность первого занятого периода.

Пусть T_1 — время первого запроса на сервер. Благодаря беспамятству мы можем сказать, что X_{t+T_1} задают новый процесс Бернулли с тем же параметром p . В этом процессе первый ноль появится во время $T_2 \sim \text{Geom}(1 - p)$. При этом T_2 в таком случае и будет равно длине занятого периода.

Время k -ого успеха. Пусть Y_k — время, в которое на сервер прибыл k -ый по счету запрос. Пусть $T_k = Y_k - Y_{k-1}$, то есть интервал между $k - 1$ -ым и k -ым запросами. При этом $Y_k = T_1 + \dots + T_k$.

Благодаря беспамятству мы можем сказать, что в каждый момент Y_k мы запускаем процесс заново. Поэтому каждый $T_k \sim \text{Geom}(p)$. Отсюда мы сразу можем сказать про Y_k , что

- $E[Y_k] = \sum_{i=1}^k E[T_i] = \frac{k}{p}$
- $\text{Var}(Y_k) = \sum_{i=1}^k \text{Var}(T_i) = \frac{k(1-p)}{p^2}$.

Также уже было в разборе контрольной, что функция распределения имеет следующий вид:

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k},$$

так как мы точно знаем, что во время t произошел успех, и нам надо расставить остальные $k - 1$ успех по оставшимся $t - 1$ ячейкам времени.

2 Слияние и разделение процесса Бернулли

Рассмотрим случай, когда у нас на сервер могут поступать запросы с двух источников. Поступление запросов от каждого из источников описывается процессом Бернулли, причем эти два процесса (назовем их X_t и Y_t) независимы. Рассмотрим процесс Z_t , где Z_t — индикаторная с.в. события “в момент времени t поступил запрос от хотя бы одного источника”. То есть $Z_t = \max\{X_t, Y_t\}$. Заметим, что Z_t не зависит от любого $Z_{s \neq t}$, так как является функцией от двух величин X_t и Y_t , не зависящих от любых других $X_{s \neq t}$ и $Y_{s \neq t}$. А также

$$\Pr[Z_t = 1] = \Pr[X_t = 1 \cup Y_t = 1] = 1 - \Pr[X_t = 0 \cap Y_t = 0] = p + q - pq,$$

где p и q — параметры процессов X_t и Y_t соответственно. Таким образом, Z_t также является процессом Бернулли с параметром $p + q - pq$. Заметьте, что вместо логического “И” мы можем применить любую другую операцию к двум поступающим запросам. Например, одновременный запрос от двух источников мы можем считать коллизией, и не считать его за запрос. Тогда Z_t также будет процессом Бернулли, но с другим параметром $p + q - 2pq$.

Мы также можем разделять процесс Бернулли на два. Пусть у нас есть процесс Бернулли Z_t с параметром p , описывающий запросы на сервер. Каждый раз, когда на сервер поступает запрос, он решает, в какой из двух выходов его отправить дальше, причем он отправляет его в первый выход с вероятностью q и во второй — с вероятностью $1 - q$, принимая это решение независимо от всего остального. В таком случае запросы, отправленные в первый выход, X_t , являются процессом Бернулли с параметром pq . Очевидно, что X_t не зависит от любого $X_{s \neq t}$, так как является функцией от величины Z_t , которая, в свою очередь, не зависит ни от каких $Z_{s \neq t}$. Аналогично, процессом Бернулли являются и запросы на втором выходе Y_t , но с параметром $p(1 - q)$. Однако процессы X_t и Y_t не являются независимыми, так как если мы знаем, что $X_t = 1$, то мы точно знаем, что $Y_t = 0$.

3 Аппроксимация процесса Бернулли

Рассмотрим процесс Бернулли с параметром p . Чему равна вероятность того, что за время n будет k успехов? По тому, что мы считали, это будет

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

Но что будет, если мы раздробим время на меньше интервалы (при этом вероятность успеха в каждый интервал будет падать обратно пропорционально длине интервала)? То есть для каждого n мы будем брать $p = \lambda/n$ для какой-то фиксированной λ и устремим $n \rightarrow \infty$. В таком случае будет:

$$\begin{aligned} p_X(k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \frac{n(n-1) \dots (n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\rightarrow 1 \cdot 1 \cdot \dots \cdot 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1 = \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

И это подводит нас к процессу Пуассона, который мы изучим на следующей паре.