Вероятности и статистика с R

Асен Чорбаджиев

November 4, 2017

1 Дискретни случайни променливи

1.1 Определение

Дискретна случайна променлива се характеризира в извадково пространство $S_x = \{u_1, u_2, ..., u_n\}$. Тогава за всяка случайна променлива съществува функция (probability mass function (PMF)) $f_X : S \to [0, 1]$:

$$f_X(x) = P(X = x), x \in S_X \tag{1}$$

за която важат следните правила:

- $f_X(x) > 0$ за всяко $x \in S$
- $\sum_{X \in S} f_X(x) = 1$
- $P(X \in A) = \sum_{X \in A} f_X(x)$ за всяко $A \subset C$

Пример: Хвърляне на монета 3 пъти с възможен изход от всяко хвърляне ези(Н) или тура(Т). Тогава:
$S = \{HHH; HTH; THH; TTH; HHT; HTT; THT; TTT\}$
и ако приемем за успех получаването на ези, например, то $S_X = \{0, 1, 2, 3\}$.
Тогава за получаване на едно ези $f_X(1) = 3/8$.

1.2 Очакване и дисперсия

Важни числови измерители са математическото очакване μ и дисперия σ^2:

$$\mu = E(x) = \sum_{x \in X} x f_X(x) \tag{2}$$

$$\sigma^2 = \sum_{X \in S} (x - \mu)^2 f_X(x) \tag{3}$$

като резултата σ се нарича стандартно отклонение.

2 Дискретни тегления със заместване

2.1 Биномно разпределение

Схема на Бернули Под схема на Бернули (n,p) се разбира серия от n независими опити, като при всеки опит има единствено два възможни враимноизключващи се изхода с вероятности съответно p и 1-p. Вероятността за постигане на k успеха е равна на:

$$P_n(k) = C_n^k p^k (1-p)^{n-k} \tag{4}$$

Когато случайната величина x е дефинирана с $P(x = k) = P_n(k)$ за всяко $x = 0, 1, ...,$ то тя е Биномно разпределена Bi(n,p). Разпределението Bi(1,p) отговаря на опит на Бернули.Математическото очакване и дисперсията са равни на $\mu = np$ и $\sigma^2 = np(1-p)$
Функциите в R за работа с комбинации е choose() и Биномно разпределение съдържат binom() в своето име.

- Разпределение dbinom()
  x <- seq(0,50,by=1)
  y <- dbinom(x,50,0.2)
  plot(x,y)
  y <- dbinom(x,50,0.5)
  plot(x,y)
- Кумулативно разпределение P($\xi \leq k$)=sum(dbinom(c(1:24), 50,1/2)):
  pbinom():
  pbinom(24,50,0.5)
  pbinom(25,50,0.5)
- Намиране на случайната величина от разпределението - qbinom():
  qbinom(0.4438624,50,1/2)
  qbinom(0.5561376,50,0.5)

- Генериране на биномно разпределени случайни величини rbinom():
rbinom(5,10,.2)
rbinom(5,100,.2)

- Когато параметърът lower.tail е равен на TRUE (default), се пресмятат вероятностите P[X $\leq$ x], обратно, P[X $>$ x].

2.2 Отрицателно Биномно разпределение и Геометрично разпределение

Когато в даден експеримент се правят тегления с двоен изход до фиксиран брой успехи r, ако има преди това има r-1 успеха и x неуспеха се използва Отрицателно Биномно Разпределение:

$$P_n(k) = C_{x+r-1}^{r-1} p^r (1-p)^x \tag{5}$$

Математическото очакване и дисперсията са равни на $\mu = pr/(1-p)$ и $\sigma^2 = pr/(1-p)^2$
Функциите в R са сходни с тези за Биномно, но залавията им се различават, като вместо binom стават nbinom, например qnbinom().
Геометричното разпределение е частен случай, когато r=1.

3 Дискретни тегления без заместване

3.1 Хипергеометрично разпределение

Двойчен експеримент, когато имаме избор без заместване, равномерно разпределена ненареденaта поредица. Нека направим избор между n-"good" и m-"bad" опции из между всички m+n възможности. Изпираме N проби и нека $x_i = 1$ е успешен избор и 0 обратното. Тогава сумата на успехи x е:

$$x = \sum_{i=1}^{N} x_i \tag{6}$$

и вероятността за i успеха е равна на:

$$P(Y = y) = \frac{\binom{n}{i}\binom{m}{N-i}}{\binom{n+m}{N}} \tag{7}$$

Математическото очакване и дисперсията са равни на $\mu = nN/(n+m)$ и $\sigma^2 = mnN(m+n-N)/((m+n)^2(m+n-1))$

Функциите в R за хипергеометрично разпределение са с име hyper със съответно d,p,q,r - за функции на плътност, разпределение, квантил и случайно разпределение.
Пример: За дадени n=N=6, r=36, m=r-N=30. Тогава i е равно на:
success=c(0:6)
dhyper(success,6,30,6)

3.2 Поасоново разпределение

Нека има опит с Биномно разпределение, което има много малка вероятност, например $p \leq 0.1$, $n \to \infty$ и $np = \lambda = const$. Тогава имаме поасоново разпределение с вероятност:

$$P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{8}$$

Основно приложение на Поасоново разпределение е при броящи процеси за поток от събития, които пристигат в случаен момент от време е да има стационарност, което означава вероятността на поява на k-брой събития във времеви интервал t зависи само от k и дължината от t. Средната стойност за поява на k-брой събития за период t е равна на λt, нарича се интензивност на потока и вероятността е равна на:

$$P_t(k) = \frac{\lambda t^k e^{-\lambda t}}{k!} \tag{9}$$

Математическото очакване и дисперсията са равни на λ, според формула (8)
Функциите в R за поасоново разпределение са с име pois(), със съответно d,p,q,r - за функции на плътност, разпределение, квантил и случайно разпределение.
Пример: Ако 12 автомобила пресичат един мост средно за минута, то средната вероятност техният брой да бъде 16 или по-малко за една минута е:
ppois(16, lambda=12)

3.3 Odds

Odds е събитие при което броят на успешни събития е разделено на броя на неслучването им:

$$Odd = \frac{p}{1-p} \tag{10}$$

Когато трябва да се оценява "релативен риск" като се сравняват две различни извадки с измервания (Таблица 1) оценките се правят с odd ratio.

Table 1:

Factor	True	False	Total
Yes	n_{11}	n_{12}	$n_{11} + n_{12}$
No	n_{21}	n_{22}	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{11}$$

Когато резултатите са само в проценти на успех от бинарни тестове, то OR се пресмята с допълване на False колоната на таблица 1 със стойността на опита до 1.

4 Упражнения:

- **Зад. 1.** Тесте карти се разбърква. Теглят се последователно от стоящата на върха на тестето карти една по една до получаване на първо Aco. Пресметнете вероятността това да бъде 5-та карта.
- **Зад. 2.**: Един пушач винаги носи в джоба си две кутии кибрит. Когато иска да запали той взима клечка от случайно избрана от двете кутии. След известно време той установява, че едната кутия е празна. Каква е вероятността в този момент другата кутия да има 21 клечки, ако в началото и в двете кутии е имало по 50 клечки? Да се реши с комбинации и разпределение.
- **Зад. 3.**: Пресметнете вероятностите за хвърляне на симетрична монета до получаване на точно 2 ези при 3,4,5,6, 7 и повече брой опити. Сравнете същите вероятности за получаване на 2 ези от фиксиран брой от 3,4,5,6,7 и повече опита за фалшива монета с вероятност за ези от 2/5. Пресметнете с odds необходимите коефициенти за залог да не загубите. Пресметнете Odds Ratio (OR) за сравняване на двата опита.
- **Зад. 4.**: Боб е баскетболен състезател с успеваемост от 70% за стрелба от тройката. Да се намери общият брой хвърляния за получаване на 3 успешни тройки с вероятност 0.18522.
- **Зад. 5.**: Социологична извадка от 100 души представлява група от 600000 гласували за опозиционна партия на последните избори. Характерно за тази партия е, че тя доминира предимно сред мъжете, като делът на жените е само 40% от всички гласували за партията.

Каква е вероятността броят на жените в извадката да е 35 или по-малко?
Нека с pop <- rep(c(0,1),c(360000, 240000)) се симулира популацията. След това да се генерира случаен вектор от 1000 елемента с резултатите за общ брой жени във всяка случайна извадка с помощта sample(pop,100)). Сравнете емпиричната честота и получения в началото на задачата теоретичен резултат.

- **Зад. 6.**: Средният брой обаждания в един колцентър е 2 на минута. Да се пресметнат вероятностите за това, че за 5 минути:
  1. Ще постъпят точно 2 обаждания
  2. Ще постъпят по-малко от 2 обаждания
  3. Не по-малко то 2 обаждания.