Lista de Exercícios - V.A.s Discretas

Renato Assunção - DCC, UFMG

2017

Esta lista de exercícios visa ao aprendizado de algumas das características das principais distribuições de probabilidade discretas. Vamos aprender um poucos sobre as seguintes distribuições: binomial, Poisson, geométrica, Pareto-Zipf.

O *R* possui um conjunto de funções para trabalhar com as principais distribuições de probabilidade. Todas operam com uma sintaxe similar. O primeiro caracter do nome da função identifica o que você quer fazer com ela: gerar números aleatórios, calcular uma probabilidade, uma probabilidade acumulada ou um quantil. Os caracteres seguinte identificam a distribuição.

Por exemplo, se quisermos trabalhar com a distribuição binomial com $n = 10$ repetiçõeos e probabilidade de sucesso $\theta = 0.15$ podemos usar:

- `rbinom(13, 20, 0.15)`: gera um conjunto de 13 inteiros aleatórios, cada um deles seguindo uma binomial $\text{Bin}(n = 20, theta = 0.15)$.
- `dbinom(13, 20, 0.15)`: se $X \sim \text{Bin}(20, 015)$, este comando calcula a função de probabilidade $\mathbb{P}(X = 13) = p(13)$ para as v.a's discretas. Podemos passar vetores como argumento. Por exemplo, `dbinom(c(10, 11, 12), 20, 0.15)` retorna o vetor $(\mathbb{P}(X = 10), \mathbb{P}(X = 11), \mathbb{P}(X = 12))$.
- `pbinom(13, 20, 0.15)`: Calcula a função de probabilidade acumulada $\mathbb{F}$ no ponto 13. Isto é, calcula $\mathbb{F}(13) = \mathbb{P}(X \leq 13)$ onde $X \sim \text{Bin}(20, 015)$.
- `pbinom(0.20, 20, 0.15)`: Calcula o quantil x associado com a de probabilidade acumulada 0.20. Isto é, calcula o valor de x tal que $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0.20$. Como X é uma v.a. discreta que acumula probabilidades aos saltos, a probabilidade acumulada até x pode ser apenas aproximadamente igual a 0.20.

Para uma Poisson, são as seguintes: `rpois`, `dpois`, `ppois` e `qpois`.

As funções correspondentes para uma gaussiana são `rnorm`, `dnorm`, `pnorm`, `qnorm`. Se quisermos trabalhar com uma gaussiana $N(\mu, \sigma^2)$, com valor esperado $\mu = 10$ e $sigma = 2$:

- `rnorm(100, 10, 2)`: gera um conjunto de 100 valores aleatórios independentes de uma v.a. $X \sim N(10, 2^2)$.
- `dnorm(11.25, 10, 2)`: retorna o valor da densidade $f(x)$ de $N(10, 2)$ no ponto $x = 11.25$. Isto é, retorna $f(11.25)$. O comando `dnorm(c(11.25, 13.15), 10, 2)` retona um vetor com os valores $(f(11.25), f(13.15))$.
- `pnorm(11.25, 10, 2)`: Calcula a função de probabilidade acumulada no ponto 11.25. Isto é, calcula $\mathbb{F}(11.25) = \mathbb{P}(X \leq 11.25)$ onde $X \sim N(10, 2^2)$.
- `pnorm(0.20, 10, 2)`: Calcula o quantil x tal que $\mathbb{F}(x) = \mathbb{P}(X \leq x) = 0.20$. Como X é uma v.a. contínua que acumula probabilidades continuamente, a probabilidade acumulada até x é exatamente igual a 0.20.

Para a exponencial, temos `rexp`, `dexp`, `pexp` e `qexp`. Para conhecer todas as distribuições disponíveis no R, digite `?distributions` ou, equivalentemente, `help(distributions)`.

---

1. Seja $X \sim$ Bin$(10, 0.4)$. Para obter e plotar (veja Figura **??**) os valores da função de probabilidade $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ uso os seguintes comandos:

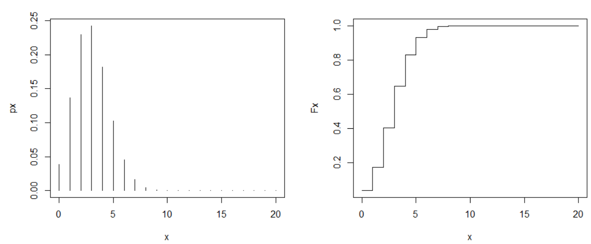

Figura 1: Função de probabilidade $\mathbb{P}(X = k)$ (esquerda) e da função de probabilidade acumulada $\mathbb{F}(x)$ (direita) de uma v.a.binomial Bin$(n = 10, \theta = 0.40)$.

```
x <- 0:10
px <- dbinom(x, 10, 0.40)
par(mfrow=c(1,2)) # janela grafica com uma linha de 2 plots
plot(x, px, type = "h") # para usar linhas verticais at\'{e} os pontos (x,px)
Fx <- pbinom(x, 10, 0.35)
plot(x, Fx, type = "s") # o argumento "s"
```

- Sua vez agora. Obtenha o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim$ Bin$(n = 20, \theta = 0.15)$. Em seguida, responda às questões abaixo.
- Qual o valor k em que $\mathbb{P}(X = k)$ é máxima? Quanto é esta probabilidade máxima?
- VISUALMENTE, obtenha uma faixa de valores (a, b) na qual a probabilidade de $X \in (a, b)$ seja próxima de 1. Procure grosseiramente obter a faixa mais estreita possível.
- O valor (teórico) de $\mathbb{E}(X)$ no caso de uma binomial é $n\theta$. Como é o comportamento da função $\mathbb{P}(X = k)$ no entorno deste valor $\mathbb{E}(X)$? Ela tem valores $\mathbb{P}(X = k)$ relativamente altos?
- Confirme esta impressão calculando $\mathbb{P}(a \leq X \leq b)$ usando a função `dnorm` ou `pnorm` do R. Por exemplo, se eu quiser $\mathbb{P}(5 \leq X \leq 8)$, uso `sum(dnorm(5:8, 20, 0.15)` ou então `pbinom(8, 20, 0.15) - pbinom(5-0.01, 20, 0.15)`. Porque eu subtraio 0.01 de 5 na chamada da segunda função?
- Use `qbinom` para obter o inteiro k tal que $\mathbb{F}(k) = \mathbb{P}(X \leq k) \approx 0.95$.
- Verifique o valor da probabilidade acumulada exata $\mathbb{F}(k)$ obtida com o inteiro acima usando `pbinom`.
- Gere 1000 valores aleatórios independentes de $X \sim$ Bin$(n = 20, \theta = 0.15)$. Estes valores cairam, em sua maioria, na faixa que você escolheu mais acima? Qual a porcentagem de valores que caiu na faixa que você escolheu?
- Compare os valores das probabilidades $\mathbb{P}(X = k)$ para $k = 0, \ldots 6$ e as frequências relativas destes inteiros nos 100 valores simulados.São parecidos?

---

2. Este problema é similar ao anterior, usando agora a distribuição de Poisson.

   - Obtenha o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim$ Poisson(λ) usando dois valores: $\lambda = 0.73$ e $\lambda = 10$.
   - O valor k em que $\mathbb{P}(X = k)$ é máximo é próximo de $\mathbb{E}(X) = \lambda$?
   - Obtenha um intervalo de valores (a, b), o mais curto possível gosseiramente, para o qual $\mathbb{P}(X \in (a, b)) \approx 1$.
   - Usando `ppois` do R, calcule $\mathbb{P}(a \leq X \leq b)$.
   - Gere 200 valores aleatórios independentes de $X \sim$ Poisson(λ) com os dois valores acima para λ.
   - Compare os valores das probabilidades $\mathbb{P}(X = k)$ para $k = 0, \ldots 6$ e as frequências relativas destes inteiros nos 100 valores simulados. São parecidos?

---

3. Este problema é similar ao anterior, usando agora a distribuição discreta de Pareto, também chamada de distribuição de Zipf. Ver `http://en.wikipedia.org/wiki/Zipf'slaw`. A distribuição de Pareto (discreta ou contínua) não está disponível em R a não ser em alguns pacotes especializados. Entretanto, não é necessário usar estes pacotes já que ela é facilmente simulada ou calculada. Veremos técnicas de simulação Monte Carlo em breve, então apenas aceite por enquanto o algoritmo abaixo.

   A distribuição discreta de Pareto possui suporte igual a $\{1, 2, \ldots, N\}$ onde N pode ser infinito. Além de N, ela possui um outro parâmetro, $\alpha > 0$. A função massa de probabilidade é dada por

$$\mathbb{P}(X = k) = \frac{C}{k^{1+\alpha}}$$

   onde C é uma constante escolhida para que as probabilidades somem 1. Observe que C é dada por

$$\frac{1}{C} = \sum_{k=1}^{N} \frac{1}{k^{1+\alpha}}$$

   Se N for um número finito, não existe uma expressão analítica para esta soma e ela deve ser calculada somando-se os valores. Se N for infinito, a expressão acima é chamada de função ζ (pronuncia-se "zeta") de Riemann:

$$\zeta(1+\alpha) = \sum_{k=1}^{\infty} \frac{1}{k^{1+\alpha}} = \frac{1}{\Gamma(1+\alpha)} \int_0^{\infty} \frac{x^\alpha}{e^x - 1} \mathrm{d}x \tag{1}$$

   (ver `http://en.wikipedia.org/wiki/Riemann_zeta_function`).

   Para alguns valores específicos de α, a função zeta $\zeta(1+\alpha)$ tem valores conhecidos exatamente. Por exemplo, para $\alpha = 1$ é possível mostrar que

$$\zeta(2) = \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots = \frac{\pi^2}{6} \approx 1.645$$

   Exceto nestes casos particulares, no caso de $N = \infty$, a constante $C = 1/\zeta(1+\alpha)$ deve ser aproximada numericamente somando-se um número grande de termos da série ou calculando numericamente a integral em (**??**). Por exemplo, para $\alpha = 1/2$, temos $\zeta(1 + 1/2) \approx 2.612$, e para $\alpha = 2$, temos $\zeta(1 + 2) \approx 1.202$.

   Tendo um valor para a constante C, podemos plotar os valores de $\mathbb{P}(X = k)$ e também da função de probabilidade acumulada $\mathbb{F}(k)$ já que

$$\mathbb{F}(k) = \mathbb{P}(X \leq k) = \sum_{i=1}^{k} \mathbb{P}(X = i) = C \sum_{i=1}^{k} \frac{1}{i^{1+\alpha}} .$$

- Usando os valores $\alpha = 1/2, 1, 2$, obtenha em R o gráfico das probabilidades $\mathbb{P}(X = k)$ e da função de probabilidade acumulada $\mathbb{F}(x)$ para uma v.a. $X \sim$ Zipf(α) com $N = \infty$. Em R, não chame a constante de integração de `c` pois este é o nome da função de concatenação de vetores e, como um defeito do R, ele não avisa que você está sobrepondo uma função-base crucial. Faça a escala horizontal variar nos inteiros de 1 a 20. Obtenha $\mathbb{F}(x)$ usando o comando `cumsum` que retorna o vetor de somas acumuladas de um vetor.
- Pelo gráfico, as probabilidades parecem cair rápido, talvez exponencialmente. Mas isto não é verdade. O comportamento dessa queda quando k aumenta é a principal razão propriedade que faz com que a distribuição power-law de Pareto (ou Zipf) seja tão importante na prática da análise de dados. Para entender como as probabilidades diminuem em direçâo a zero a medida que k cresce, obtenha a raz ao entre valores sucessivos de $\mathbb{P}(X = k)$. Isto é, mostre que
$$\frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = \left(\frac{k}{k+1}\right)^{1+\alpha}$$
Perceba agora que, quando k cresce, $k/(k+1)$ é sempre menor que 1 mas cada vez mais próximo de 1 e portanto
$$\mathbb{P}(X = k+1) \approx \mathbb{P}(X = k)$$
se k for bem grande. As duas probabilidades serão pequenas mas quase idênticas. Isto é, a medida que k cresce, as probabilidades decaem muito lentamente, quase nadaquando $k$for bem grande.
- Quando $\alpha > 0$ crescer, o que você esperar acontecer ao gerar inteiros Zipf com estes α grandes em relação à geração com α apenas ligeiramente maior que 1.
- Faạa um gráfico dos pontos $(\log(k), \log(\mathbb{P}(X = k)$. O resultado é o que você esperava? Usando `abline(log(C), -(1 + alpha))`, sobreponha uma reta com intercepto $log(C)$ e inclinaç ao $-(1 + \alpha)$.
- Chega de análise teórica, vamos simular por MOnte Carlo alguns valores Zipf agora. A função R abaixo faz isto para você:

```
rzipf = function(nsim = 1, alpha = 1, Cte = 1/1.645)
{
 res = numeric(nsim)
 for(i in 1:nsim){
   x = -1
   k = 1
   F = p =  Cte
   U = runif(1)
   while( x == -1){
     if(U < F) x = k
     else{
       p = p * (k/(k+1))^(1+alpha)
       F = F + p
       k = k+1
     }
   }
 res[i] = x
 }
 res
}
```

Por default, a função assume $\alpha = 1$ e fornece também a constante C. Para gerar $nsim = 400$ valores com estes argumentos default, basta digitar `rzipf(400)`. Para gerar 400 valores de uma Zipf com $\alpha = 1/2$ e com a constante $C = 1/2.612$ determinada por este valor de α, basta digitar `rzipf(400, 1/2, 1/2.62)`.

Agora, a tarefa: gere 400 valores de Zipf com $\alpha = 1/2, 1, 2$ (as constantes estão no texto acima). Verifique que apesar da maioria dos valores ficar num intervalo limitado, valores extremamente grandes (relativamente aos demais) são gerados com facilidade. Repita a geração algumas vezes para observar este efeito. Reporte na lista apenas uma dessas repetições.