Estatística Computacional
Lista de Problemas

Mauro Campos
Departamento de Estatística, Universidade Federal do Espírito Santo*

19 de setembro de 2018

Problema 1. O problema da ruína do jogador (the gambler's ruin problem) pode ser apresentado como segue. Considere que dois jogadores, denotados por J_1 e J_2, possuem um capital inicial de a e b unidades monetárias respectivamente, onde a e b são inteiros positivos e $d = a+b$. Suponha que esses jogadores estão realizando uma sequência de apostas um contra o outro e que a cada aposta, o capital de J_1 aumenta ou diminui de 1 unidade monetária dependendo se ele ganha ou perde a aposta para J_2. Observe que o capital total dos jogadores a cada aposta é sempre igual a d. Apostas são realizadas até que um dos jogadores perde todo seu capital (e consequentemente, seu oponente alcance o capital total de d unidades monetárias). O jogador J_1 ganha uma aposta de J_2 com probabilidade p ($0 < p < 1$) e perde de J_2 com probabilidade $q = 1-p$ independente da aposta que está sendo realizada. Descreva esse processo como uma cadeia de Markov, onde X_n representa o Capital de J_1 no tempo n.

- Calcule $P_a^{d,p} = \Pr(\text{Ruína de } J_1 | X_0 = a)$, onde "Ruína de J_1"é o evento $\{X_n = 0 : \text{para algun } n > 0\} = \cup_{n>0}[X_n = 0]$.
- Calcule $\mu_a^{d,p} = \mathrm{E}(T|X_0 = a)$, onde T é tempo de espera até que um dos jogadores chegue à ruína.
- A ruína de um dos jogadores é certa?
- Desenvolva um algoritmo para calcular valores que completam as Tabelas 1 e 2.
- Suponha que J_2 é um banco (com capital muito grande). Nesse caso, calcule a probabilidade de ruína de J_1 e o tempo médio de espera até a ocorrência desse evento.

Tabela 1: Valores para $P_a^{20,p}$.

a	$p = 1/4$	$p = 1/2$	$p = 3/4$
0	1	1	1
1			
$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	0	0	0

Tabela 2: Valores para $\mu_a^{20,p}$.

a	$p = 1/4$	$p = 1/2$	$p = 3/4$
0	0	0	0
1			
$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	0	0	0

Problema 2. Desenvolva um algoritmo para simular um processo de Poisson com taxa $\lambda > 0$.

Problema 3. Um baralho com N cartas é embaralhado. Dizemos que um encontro ocorre quando a carta i se encontra na i-ésima posição após o embaralhamento. Escreva um algoritmo para estimar a probabilidade de ocorrer pelo menos um encontro após o embaralhamento. Desenvolva um estudo de simulação e verifique o que acontece com o valor dessa probabilidade quando N cresce. Encontre respostas exatas para esse problema e compare seus resultados teóricos com suas estimativas via simulação.

*DEST, UFES. Av. Fernando Ferrari 514, 29075-910, Vitória ES.

Problema 4. Estude a distribuição normal multivariada. Utilize o método da decomposição espectral para simular 200 valores de $\mathbf{X} = (X_1, X_2)^T \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ onde

$$\boldsymbol{\mu} = (1,1)^T \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix}.$$

Apresente um gráfico bidimensional onde se pode ver (ao mesmo tempo) as curvas de nível da densidade de $\mathbf{X}$ e os valores simulados.

Tabela 3: Amostra aleatória simples (observada) de $X \sim \mathrm{Ga}(a,b)$.

2.200	4.272	2.637	2.050	3.721	2.743	2.871	1.991	3.965	2.224
2.046	2.673	3.759	3.973	4.177	2.123	1.969	2.609	4.079	2.142
1.913	1.702	4.040	3.835	1.103	2.249	0.416	3.191	2.035	0.897
2.131	2.781	3.059	2.916	3.862	3.778	1.419	4.889	2.614	1.748

Problema 5. Assuma que os dados da Tabela 3 formam uma amostra aleatória simples da distribuição de X, onde $X \sim \mathrm{Ga}(a,b)$. Lembre-se que a densidade de X é dada por

$$\mathrm{Ga}(x|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & \text{se } x > 0 \ (a,b > 0) \\ 0 & \text{caso contrário.} \end{cases}$$

e que $\mathrm{E}(X) = a/b$ e $\mathrm{Var}(X) = a/b^2$.

- Encontre formas fechadas para estimadores de a e b via métodos dos momentos.
- Utilize o método de Newton-Raphson para encontrar estimativas de máxima verossimilhança (MV) para a e b. Use as estimativas via métodos do momento como ponto de partida.
- Use um simulated annealing para encontrar estimativas de MV para a e b.
- Use um algoritmo genético para encontrar estimativas de MV para a e b.
- Desenvolva um estudo de simulação para compara o desempenho dos métodos Newton-Raphson, simulated annealing e algoritmo genético para encontrar estimativas de MV para a e b.

Problema 6. Resolva o Problema 5 usando a função `optim()` do R.

```
> LogLik <- function(theta,sumx,sumlogx,n){
+   a <- theta[1]; b <- theta[2]
+   out <- (-1)*(n*a*log(b) - n*log(gamma(a)) - sumx*b + sumlogx*(a-1))
+   return(out)
+ }
> m1 <- mean(dados)
> m2 <- mean(dados^(2))
> a.mm <- m1^(2)/(m2-m1^(2)); b.mm <- m1/(m2-m1^(2)) # metodo dos momentos
> result <- optim(c(a.mm,b.mm),LogLik,method="L-BFGS-B",lower=c(0,0),upper=c(Inf,Inf),
+                 sumx=sum(dados),sumlogx=sum(log(dados)),n=length(dados))
```

Problema 7. Um problema de programação linear (PPL) é um problema de otimização onde tanto a função objetivo quanto as restrições são funções lineares das variáveis de decisão do problema. Sem perda de generalidade, a forma padrão de um PPL é definida como:

$$\begin{aligned} \min \quad & \mathbf{c}^T\mathbf{x} \\ \text{sujeito a} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

onde $\mathbf{x} = (x_1, \ldots, x_n)$ é um vetor coluna n-dimensional, $\mathbf{c}^T$ é um vetor linha n-dimensional, $\mathbf{A}$ é uma matrix $m \times n$ e $\mathbf{b}$ é um vetor coluna m-dimensional. A desigualdade $\mathbf{x} \geq \mathbf{0}$ significa que cada componente de $\mathbf{x}$ é não negativa. Outras formas são também possíveis:

$$\begin{aligned} \min \quad & \mathbf{c}^T\mathbf{x} \\ \text{sujeito a} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned} \qquad\qquad \begin{aligned} \max \quad & \mathbf{c}^T\mathbf{x} \\ \text{sujeito a} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

Qualquer ponto $\mathbf{x}$ que satisfaz as restrições do problema é chamado de solução viável. Se o problema é de minimização, o principal objetivo é encontrar uma solução viável que retorna o menor valor da função objetivo. Se o problema é de maximização, o principal

objetivo é encontrar uma solução viável que retorna o maior valor da função objetivo. Um exemplo simples é como segue:

$$\begin{aligned} \min \quad & 0.3x_1 + 0.9x_2 \\ \text{sujeito a} \quad & x_1 + x_2 \geq 800 \\ & -0.21x_1 + 0.30x_2 \geq 0 \\ & 0.03x_1 - 0.01x_2 \geq 0 \\ & (x_1, x_2) \geq 0 \end{aligned}$$

No R, um PPL pode ser resolvido usando a função `solveLP()` do pacote `linprog`. Vejamos como obter a solução viável ótima para o exemplo acima. Essa solução é dada por $\mathbf{x}^* = (470.59, 329.41)$, cujo o valor objetivo ótimo é 437.65.

```
> library(linprog)
> cvec <- c(0.3,0.9)
> bvec <- c(800,0,0)
> Amat <- rbind(c(1,1),c(-0.21,0.30),c(0.03,-0.01))
> Amat
      [,1]  [,2]
[1,]  1.00  1.00
[2,] -0.21  0.30
[3,]  0.03 -0.01
> result <- solveLP(cvec,bvec,Amat,maximum=FALSE,const.dir=rep(">=",length(bvec)))
> result$solution
       1        2
470.5882 329.4118
> result$opt
[1] 437.6471
> result <- solveLP(cvec,bvec,Amat,maximum=FALSE,const.dir=rep(">=",length(bvec)),lpSolve=TRUE)
> result$solution
       1        2
470.5882 329.4118
> result$opt
[1] 437.6471
```

Estude um pouco mais a função `solveLP()` e resolva o seguinte PPL. Na terra de Oz, existe uma marca de automáveis chamada MC-Auto, que possui 3 fábricas em Oz: uma na cidade Azul, uma na cidade Vermelha e outra na cidade Verde. MC-Auto mantém duas centrais de distribuição: uma na região norte de Oz e outra região sul de Oz. A capacidade produtiva das fábricas para o próximo trimestre é de: 1000 carros na fábrica Azul, 1500 carros na fábrica Vermelha e 1200 carros na fábrica Verde. A demanda trimestral nas duas centrais de distribuição é de: 2300 para a central norte e 1400 para a central sul. Os custos de transporte por carro nas diferentes rotas que ligam as fábricas as centrais de distribuição são dados na Tabela 4. Proponha uma programação ótima de distribuição dos carros fabricados para as centrais de distribuidoras para próximo trimestre que minimize o custo total com transporte.

Tabela 4: Custo ($) de transporte por carro.

Fábrica	Região Norte	Região Sul
Azul	80	215
Vermelha	100	108
Verde	102	68

Problema 8. Considere que os dados da Tabela 5 formam uma amostra aleatória simples da distribuição de uma variável aleatória X.

Tabela 5: Amostra da distribuição de X.

1.319	2.268	0.826	1.465	0.729
0.755	3.694	0.393	4.200	0.950
0.075	1.743	3.849	5.366	2.470

- Estime o erro-padrão da média amostral, $\bar{X}$, via bootstrap não-paramétrico.
- Compare o resultado obtido via bootstrap com o valor $\hat{\text{EP}}(\bar{X}) = S/\sqrt{n}$, onde S é desvio-padrão amostral.
- Construa um intervalo de 95% de confiança (aproximadamente) para $\mu = \text{E}(X)$ via bootstrap não-paramétrico.

Problema 9. Considere os dados do Problema 8 e assuma que

$$x_i \overset{\text{iid}}{\sim} \text{Exp}(x;\beta) \begin{cases} (1/\beta)\cdot\exp(-x/\beta) & x > 0 \\ 0 & \text{caso contrário.} \end{cases} \tag{1}$$

- Encontre o estimador de máxima verossimilhança, $\hat{\beta}_{MV}$, para β.
- Encontre a distribuição amostral de $\hat{\beta}_{MV}$, sua média e seu erro-padrão.
- Estime a distribuição amostral de $\hat{\beta}_{MV}$ via bootstrap paramétrico.
- Estime o erro-padrão de $\hat{\beta}_{MV}$ via bootstrap paramétrico.
- Construa um intervalo de 95% de confiança (aproximadamente) para β via bootstrap paramétrico.
- Compare os resultados teóricos com os resultados obtidos via bootstrap.

Problema 10. Considere que os dados da Tabela 6 formam uma amostra aleatória simples da distribuição conjunta do vetor aleatório $\mathbf{X} = (X_1, X_2)$. Construa um intervalo de 95% de confiança (aproximadamente) para o parâmetro $\theta = \text{CorrelaçãoAmostral}(X_1, X_2)$ via

Tabela 6: Amostra da distribuição de $\mathbf{X} = (X, Y)$.

X_1	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594
X_2	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43	3.36	3.13	3.12	2.74	2.76	2.88	3.96

bootstrap não-paramétrico.

Problema 11. Utilize o método de integração Monte Carlo para estimar a seguinte integral:

$$\int_0^\infty \frac{x}{(1+x^2)^2} dx.$$

Problema 12. Seja $U \sim U(0,1)$. Use o método de integração Monte Carlo para estimar:

- $\text{Cov}(U, \sqrt{1-U^2})$
- $\text{Cov}(U^2, \sqrt{1-U^2})$.

Problema 13. Considere uma variável aleatória discreta X assumindo valores em $S = \{1,2,3\}$ com função de probabilidade dada por $X \sim \pi = (2/6, 3/6, 1/6)$. Nessa questão, construa explicitamente a cadeia de Markov (homogênea) do algoritmo de Metropolis-Hastings para estimar a quantidade $\theta = \text{E}(X^2)$. Para isso, considere que a função q que propõe um estado "candidato"para essa cadeia seja definida por:

$$q = (q(x,y)) = \begin{pmatrix} 1/4 & 2/4 & 1/4 \\ 4/9 & 4/9 & 1/9 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}. \tag{2}$$

- Encontre a matriz de transição P da cadeia, assumindo (obviamente) que espaço de estados é S.
- Mostre que $\pi P = \pi$. Ou seja, mostre que π é uma distribuição estacionária para a cadeia cuja matriz de transição é P. Mostre também que π é a única distribuição estacionária dessa cadeia.
- Simule explicitamente valores da distribuição estacionária π, considerando simplesmente a matriz de transição P e, com os valores simulados, estime θ. Compare o resultado obtido com o resultado exato.
- Por fim, implemente o algoritmo de Metropolis-Hastings para estimar $\theta = \text{E}(X^2)$. Compare o resultado obtido com o resultado obtido no item anterior.

Problema 14. Considere o seguinte modelo hierárquico

$$X|\theta \sim \text{Normal}(\theta, 1) \qquad \theta \sim \text{Cauchy}(0,1) \tag{3}$$

e assuma que $\mathbf{x} = (0.93, 2.94, 0.36, 3.08, 3.68, 2.36, 1.12, 3.71, 2.38, 3.14)$ representa uma amostra observada da distribuição de $X|\theta$. Implemente o algoritmo de Metropolis-Hastings para simular valores da distribuição $p(\theta|\mathbf{x})$ (a distribuição a posteriori de θ) e estimar θ.

Problema 15. Seja (X,Y) um vetor aleatório cuja distribuição conjunta é dada por

$$f(x,y) \propto \frac{n!}{x!(n-x)!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \tag{4}$$

onde $x = 0, 1, \ldots, n$ and $0 \leq y \leq 1$. Use o amostrador de Gibbs para obter estimativas para $\mu_{X,Y} = \text{E}(XY)$, $\mu_X = \text{E}(X)$, e $\mu_Y = \text{E}(Y)$.

Referências

W. Martinez & A. Martinez. (2002). Computational Statistics Handbook with MATLAB. New York: Chapman & Hall/CRC.

S. Ross. (1997). Simulation, 2nd ed. New York: Academic Press.