EP4

Flora Matsubara Hollander nUSP 11811221
Rodrigo Meireles de Oliveira nUSP 7580967

Junho 2021

Objetivo

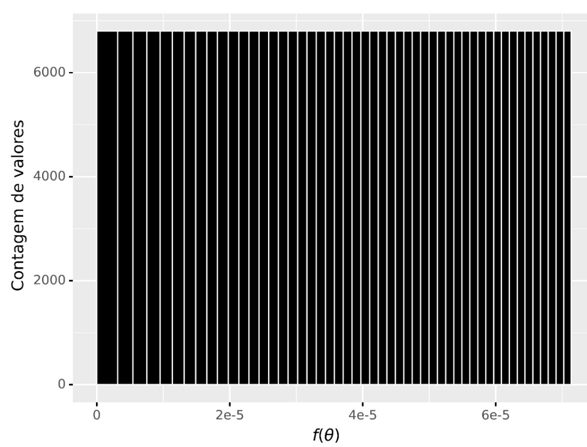
A partir do modelo estatístico apresentado em aula, neste exercício-programa implementaremos a função *U(v)* que aproxima bem a função-verdade pedida *W(v)* utilizando as definições e aproximações do enunciado. É necessário que o erro desta seja menor que 0.05%, portanto foi necessário o cálculo de n (nº de pontos θ_n) suficientemente grande.

Método

Primeiro, para a função $f(\theta|x,y)$ definimos *f(alpha, theta)* utilizando a função densidade Dirichlet. Em *im_f* armazenamos o resultado de $f(alpha, theta)$ calculado para cada um dos n thetas sorteados por Dirichlet. A partir desses valores, geramos os bins com a quantidade de valores por bin desejada com a função *qcut()* do pacote *pandas*.

```
(7.136e-05, 7.1364e-05]        34
(2.9684e-05, 2.9691e-05]       34
(2.9737e-05, 2.9745e-05]       34
(2.9729e-05, 2.9737e-05]       34
(2.9721e-05, 2.9729e-05]       34
                               ..
(5.1643e-05, 5.165e-05]        34
(5.1636e-05, 5.1643e-05]       34
(5.163e-05, 5.1636e-05]        34
(5.1623e-05, 5.163e-05]        34
(-9.9985344e-06, 5.259e-08]    34
```

Bins de $f(\theta)$ e número de valores por bin



Histograma dos bins de $f(\theta)$

Para encontrar uma aproximação de $W(v)$, temos que

$$W(t_j) - W(t_{j-1}) \approx \frac{1}{k} \Rightarrow W(t_j) \approx \frac{j}{k}$$

Ou seja, como a integral em intervalo pode ser aproximada por $1/k$, o valor até um ponto pode ser aproximada pela soma da integral de cada intervalo anterior a ele. A função $U(v)$ realiza uma busca binária nos valores dos bins para localizar o índice do intervalo aproximado de v e divide seu valor por k.

A fim de garantir que o erro cometido pela função seja menor que 0.05%, utilizamos $k = 10000$, já que queremos $\frac{1}{k} \leq 0.05\%$. Além disso, utilizamos a inequação de Dvoretzky-Kiefer-Wolfowitz[1] para estimar n suficientemente grande. Para um intervalo de confiança com significância 95%, sendo $\epsilon = 0.05\%$ o erro desejado, temos:

$$P(\underset{t \exists R^d}{sup} |F_n(t) - F(t)| > \epsilon) \leq (n+1)de^{-2n\epsilon^2} < 0.05$$

Em que $d = 3$ é a dimensão dos vetores. Segunda a referência, para n suficientemente grande podemos substituir $(n+1)$ por 2, obtemos

$$6e^{-2n0.0005^2} < 0.05 \quad \Rightarrow \quad n \geq -\frac{ln(\frac{0.05}{6})}{5 \times 10^{-7}}$$

Código

```python
'''Flora Matsubara Hollander nUSP 11811221 & Rodrigo Meireles de Oliveira nUSP 7580967
  Declaramos este trabalho como público.'''

import numpy as np
import pandas as pd
import math

rng = np.random.default_rng(2021) #seed para reproducibilidade
x = [3, 2, 1] #x e y do enunciado <==== PARÂMETROS PARA OS MONITORES
y = [1, 2, 3]
alpha = [i+j for i, j in zip(x,y)]
k = 10000
beta_cte = np.prod([math.gamma(a) for a in alpha])/math.gamma(np.sum(alpha))

def f(alpha, theta): #função densidade Dirichlet
  return np.prod([t**(a - 1) for a, t in zip(alpha, theta)])/beta_cte

n = int(np.ceil(np.log(0.05/6)/-5e-7))  #aplicando desigualdade DKW para encontrar n tal que o erro de integração é menor que 0.0005 em 95% das vezes
print(f'Precisamos de {n} pontos da Dirichlet.')

thetas = rng.dirichlet(alpha, size = n)
im_f = [f(alpha, theta) for theta in thetas]
max_f = np.max(im_f)
bins = pd.qcut(im_f, q = k, retbins = True)[1] #função do pandas que cria os bins com quantidade de valores iguais por quantil
bins[0] = 0

def U(v):
  '''Aproximação de W(v), recebe v e retorna o valor da integral das caudas i.e. a integral de f(t) para todo t tal que f(t) <= f(v)'''
  if v <= min(im_f):
    return 0
  elif v >= max(im_f):
    return 1
  else:
    return (np.searchsorted(bins, v))/k #busca binária nos valores dos intervalos, ele descobre o valor de v pra saber em qual intervalo o v fica
                                        #pq a integral é até <= v
```

O código acima retorna a função $U(v)$ como desejado. Os testes **U(1)** e **U(0.5)** retornam **0.0589** e **0.0244** respectivamente, condizente com a precisão pedida.

[1] Baseado em `https://en.wikipedia.org/wiki/Dvoretzky%E2%80%93Kiefer%E2%80%93Wolfowitz_inequality`