Simulação Discreta
Aplicação do Método Kolmogorov-Smirnov

Samuel F. Aguiar[1]

[1]Faculdade de Computação – Instituto Ciências Exatas e Naturais
Universidade Federal do Pará (UFPA)
Av. Augusto Correa 01, 66075-090 – Belém – PA – Brasil

1. Descrição

Esta tarefa consiste em aplicar o método Kolmogorov-Smirnov para uma base de dados. Os alunos devem identificar qual a função de probabilidade que se adequa aos dados observados, para um $\alpha = 0,05$.

Os alunos também precisam observar que os dados de entrada devem ser tratados adequadamente, com a remoção de outliers que possam vir a comprometer os resultados aguardados.

2. Coleta e inserção de dados

Os dados foram disponibilizados em um arquivo chamado ”entrada-trabalho-metodo-ks.txt”junto com o PDF da atividade. São eles:

27, 29, 38, 45, 46, 47, 48, 49, 49, 50, 54, 54, 54, 54, 55, 55, 55, 57, 58, 58, 59, 60, 60, 60, 61, 62, 62, 64, 65, 65, 66, 67, 67, 68, 68, 69, 69, 70, 70, 70, 70, 71, 71, 71, 71, 72, 72, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 77, 77, 78, 78, 78, 78, 78, 78, 79, 79, 79, 79, 79, 79, 79, 79, 79, 80, 80, 81, 81, 81, 81, 82, 82, 82, 83, 83, 84, 84, 84, 84, 84, 84, 85, 86, 86, 86, 87, 87, 87, 87, 87, 87, 87, 87, 87, 88, 88, 88, 88, 88, 88, 88, 88, 89, 89, 89, 89, 89, 89, 90, 90, 91, 91, 91, 91, 92, 92, 92, 93, 93, 93, 94, 94, 94, 94, 95, 95, 96, 96, 96, 97, 98, 99, 99, 99, 99, 99, 100, 100, 100, 101, 101, 101, 101, 101, 101, 102, 102, 103, 104, 104, 104, 104, 104, 104, 104, 104, 104, 105, 106, 108, 108, 108, 109, 110, 111, 111, 111, 111, 112, 113, 114, 114, 114, 114, 115, 116, 116, 117, 118, 119, 120, 122, 124, 126, 127, 127, 127, 127, 131, 134, 230, 315.

Os dados precisavam ser tratados para serem interpretados corretamente pelo Python, umas vez que o Python não reconhece vetores com espaços, mas sim com vírgulas. O seguinte programa fez essa operação:

```
caminho_do_arquivo = input("Digite aqui o caminho do arquivo:")
f = open(caminho_do_arquivo, "r")
string = f.read()

string = string.replace("\n", " ")
string = string.replace("  ", " ")

dados = string.split()

for i in range(len(dados)):
    dados[i] = int(dados[i])
```

```
print(dados)
```

3. Tratamento de dados

Utilizando Estatística Descritiva, obtem-se as seguintes medições para o conjunto de dados:

Medidas de Centralidade	-
Média	88.1633
Mediana	87
Moda	79
Mínimo	27
Máximo	315
Medidas de Dispersão	-
Amplitude	288
Desvio Padrão	27.7343
Variância	769.1920
Coeficiente de Variação	3.6056
Coeficiente de Assimetria	0.1244

Esses dados precisam ser tratados, pois apresentam outliers que podem afetar numa análise mais precisa dos dados. Para isso, devemos calcular a amplitude interquartil, que pode ser dada pela diferença entre o terceiro e o primeiro elementos do quartil, i.e. $A = Q_3 - Q_1$

Nessa lista, encontramos dois outliers, sendo ambos extremos: 230 e 315

Com a remoção dos dois outliers, agora temos os dados se apresentando da seguinte forma:

Medidas de Centralidade	-
Média	86.3200
Mediana	87
Moda	79
Mínimo	27
Máximo	134
Medidas de Dispersão	-
Amplitude	107
Desvio Padrão	20.3432
Variância	413.8468
Coeficiente de Variação	4.9156
Coeficiente de Assimetria	0.1244

4. Histograma

Agora, com os dados devidamente tratados, é necessário a criação de um histograma com o novo conjunto.

O número de classes do histograma é dado pela fórmula $K = 1 + 3,3 \log_{10} n$. Com isso, temos nove classes.

O histograma é representado da seguinte forma:

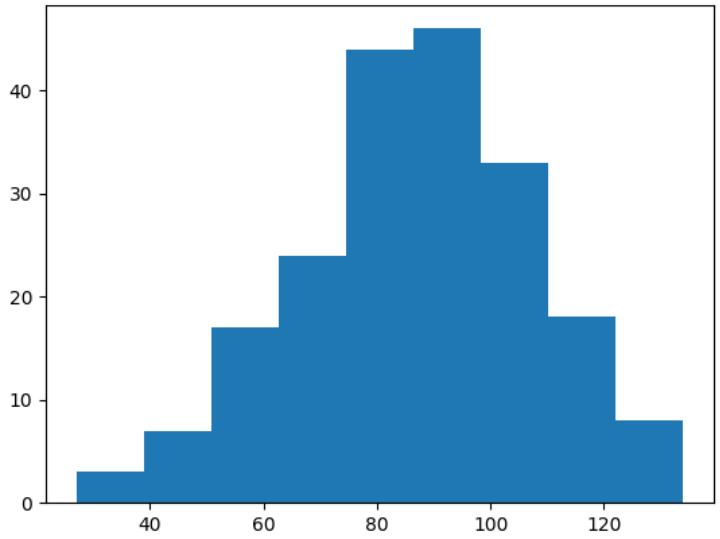

Figura 1. Histograma desenhado com os dados tratados.

As frequências desse histograma estão representadas no seguinte quadro:

Classes	Intervalos	Frequência
1	$27 < valor \leq 38,888...$	3
2	$38,888... < valor \leq 50,777...$	7
3	$50,777... < valor \leq 62,666...$	17
4	$62,777... < valor \leq 74,555...$	24
5	$74,555... < valor \leq 86,444...$	44
6	$86,444.. < valor \leq 98,333...$	46
7	$98,333... < valor \leq 110,222...$	33
8	$110,222... < valor \leq 12,111...$	18
9	$122,111... < valor \leq 134$	8

5. Aplicando o método Kolmogorov-Smirnov

Com os dados coletados e as frequências devidamente anotadas em uma planilha do excel disponível em: `https://github.com/RFLMNaguiar/trabalho-simulacao-discreta-kolmogorov-smirnov-samuel-aguiar`, podemos fazer os cálculos necessários.

Observando o histograma da figura 1, percebemos que a distribuição se aparenta ser próxima da distribuição normal ou talvez da lognormal.

Aplicando o método, a maior diferença entre os valores da distribuição normal e a distribuição obtida é de 0,037465. Para uma confiabilidade $\alpha = 0.05$, fazemos $1.36/\sqrt{200}$, que é 0,096166 (D crítico). Como a diferença entre a distribuição normal e a obtida é menor que nosso D crítico, então a distribuição é aderente ao conjunto de dados.

6. Apêndice

Link para o repositório com todos os arquivos do trabalho: `https://github.com/RFLMNaguiar/trabalho-simulacao-discreta-kolmogorov-smirnov-samuel-aguiar`