UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Relatório do trabalho de probabilidade

Alunos: Gabriel Machado e Tadeu Tupinambá

Novembro
2017

Instituto de Biociências, Letras e Ciências Exatas
Departamento de Ciências de Computação e Estatística
Programa de graduação

Relatório do trabalho de probabilidade

Relatório do trabalho da disciplina Probabilidade e Estatística.

Alunos: Gabriel Machado e Tadeu Tupinambá

Professora: Adriana Barbosa Santos

Novembro
2017

Conteúdo

1 Objetivo

Estudar modelos de probabilísticos e o Teorema do Limite Central por meio de experimentos de simulação.

2 Metodologia

Para a implementação de ambos experimentos foi usado a linguagem de programação Python pela facilidade utiliza-la e por ser bem versátil em divérsas áreas, incluindo a de estatística.

2.1 Experimento 1

No primeiro experimento foi codificado três funções:

- Uma função que vai gerar histograma de distribuições uniformes que recebe como parâmetro o número de amostras que será gerada para analisar a geração de números pseudo-aleatórios.
- Uma função para simular a variável aleatória Y, proposta no trabalho, com tamanho e alpha como parâmetros e a partir dos dados gerar um histograma.
- Uma função para simular uma variável aleatória exponencial com tamanho e alpha como parâmetros e a partir dos dados gerar um histograma.

2.2 Experimento 2

Para o segundo experimento foi codificado apenas um método de leitura de arquivos csv e alguns outros para fazer alguns cálculos de propriedades estatísticas.

Diferentemente do primeiro experimento, nesse os dados não foram gerados de maneira aleatória, foram obtidos através de uma base de dados aberta. Essa base tem os dados de quantidade de pessoas por cidade dos Estados Unidos.

Para conseguir usar os dados obtidos na base aberta foi preciso filtrar os dados nas seguintes etapas:

1. Tranformar arquivo de dados do formato xls para csv (para facilitar o manuseio).
2. Agrupar dados de cidades em estados.
3. Remover colunas de informações não desejadas
4. Extrair conhecimento dos dados através de gráficos

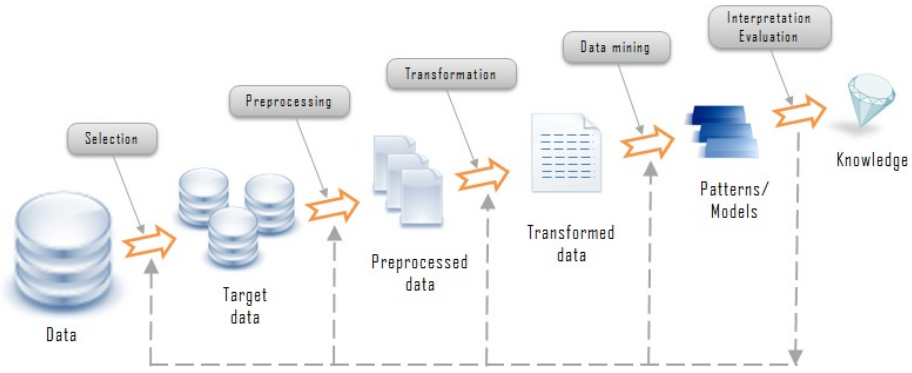

Figura 2.2.1: Processo para obtenção de informação.

Esse processo de extração e limpeza de dados podem ser observado como na figura 2.2.1.

3 Resultados

Nas seções seguintes será mostrado os testes feitos e seus resultados dos dois experimentos propostos.

3.1 Experimento 1

Os primeiros testes a serem analisados são os histogramas da distribuicão uniforme, feitos para analisar o gerador de números pseudo-aleatórios da linguágem Python.

3.1.1 Distribuição Uniforme

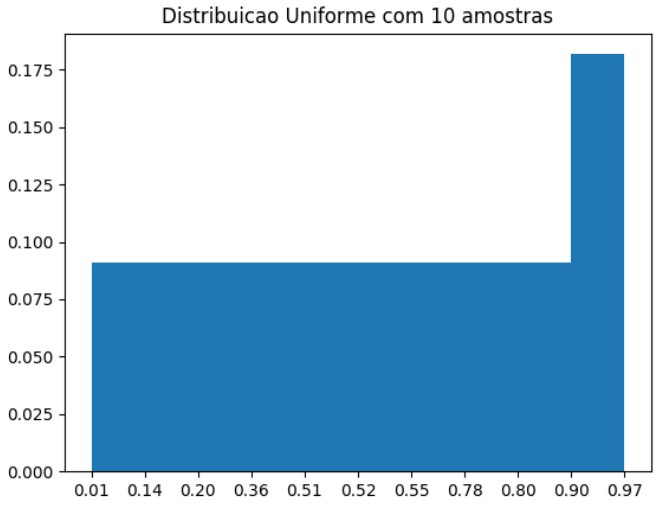

Figura 3.1.1: Histograma da distribuição uniforme com 10 amostras.

Pelo histograma apresentado na figura 3.1.1 não conseguimos obter muita informação relevante pois o número de amostras (10) é muito baixo e portanto oscilações como a observada é normal.

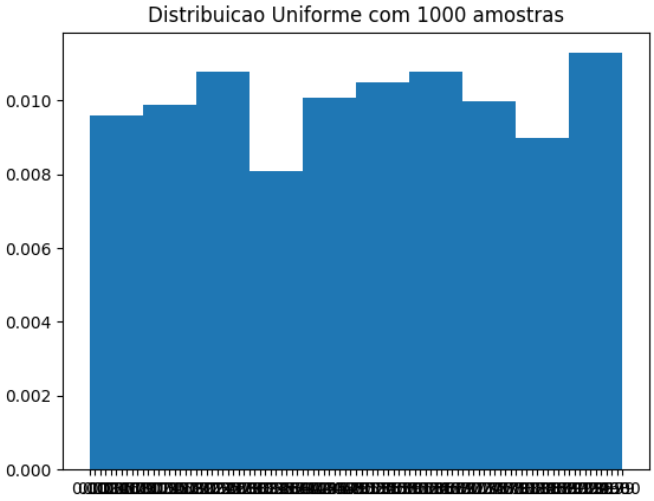

Figura 3.1.2: Histograma da distribuição uniforme com 1000 amostras.

Ao aumentar o número de amostras para 1000 conseguimos começar a notar um nivelamento um pouco mais desejável para uma sequência de dados aleatórios distribuídos uniformemente, como mostrado na figura 3.1.2.

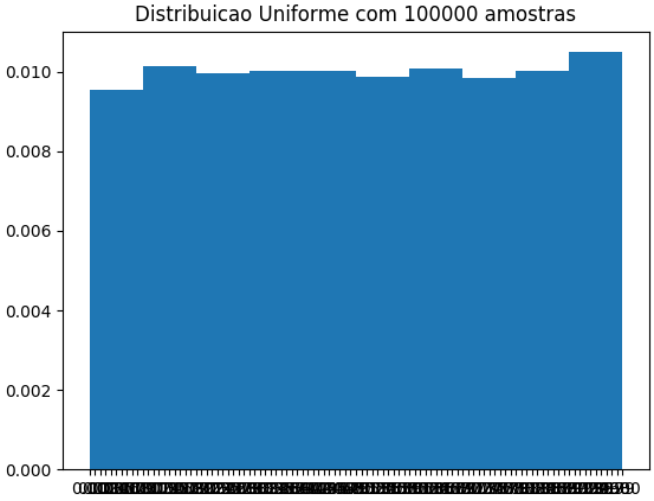

Figura 3.1.3: Histograma da distribuição uniforme com 100000 amostras.

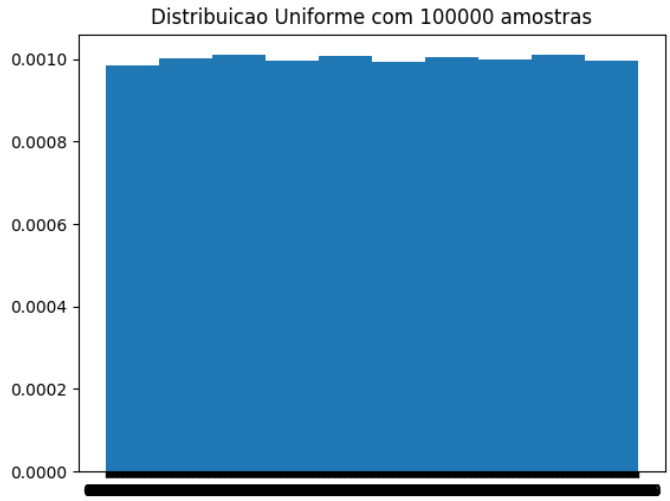

Figura 3.1.4: Histograma da distribuição uniforme com 100000 amostras com intervalos menores.

Com 100000 amostras os dados do histograma ficam ainda mais uniformes como mostrado na figura 3.1.3. Em 3.1.4 observamos 100000 amostras novamente, porém agora com intervalos menores.

3.1.2 Análise de variável aleatória Y

Nessa parte do experimento foi feito vários testes com a função de distribuição da v.a. Y e a de exponencial, comparando as duas para ver se a Y se comporta como uma exponencial.

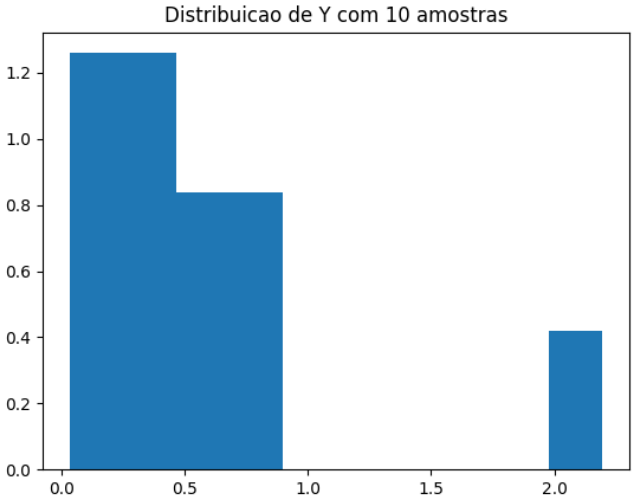

Figura 3.1.5: Histograma da distribuição da v.a. Y com 10 amostras.

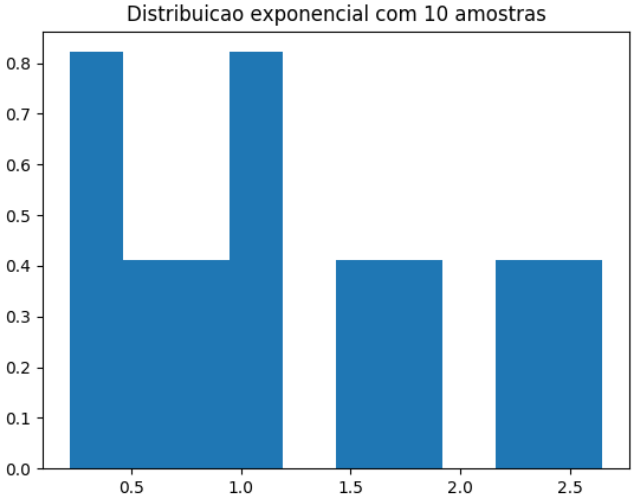

Figura 3.1.6: Histograma da distribuição exponencial com 10 amostras.

No primeiro teste com 10 amostras fica difícil comparar as duas pois nenhuma segue um padrão com poucas amostras assim como é possível ver nas figuras 3.1.5 e 3.1.6.

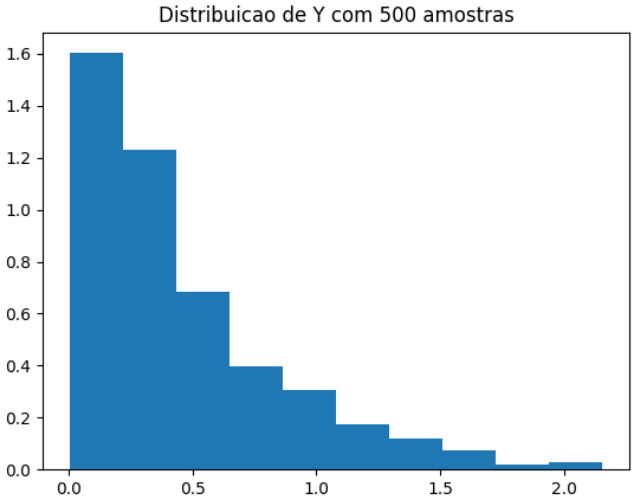

Figura 3.1.7: Histograma da distribuição da v.a. Y com 500 amostras.

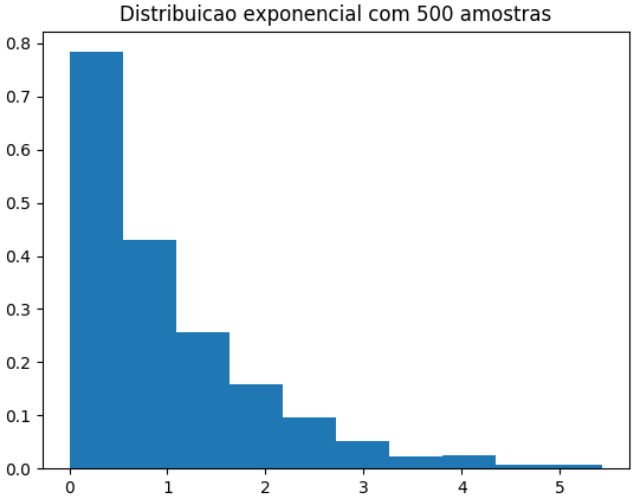

Figura 3.1.8: Histograma da distribuição exponencial com 500 amostras.

Com 500 amostras já é possível observar as semelhanças dos dois gráficos e o padrão de decaimento exponencial.

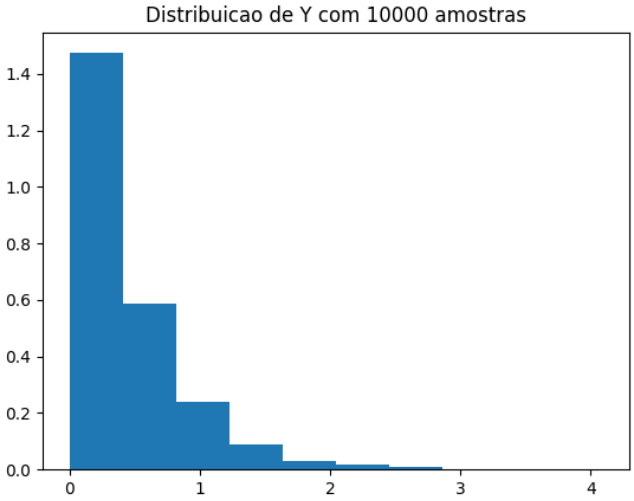

Figura 3.1.9: Histograma da distribuição da v.a. Y com 10000 amostras.

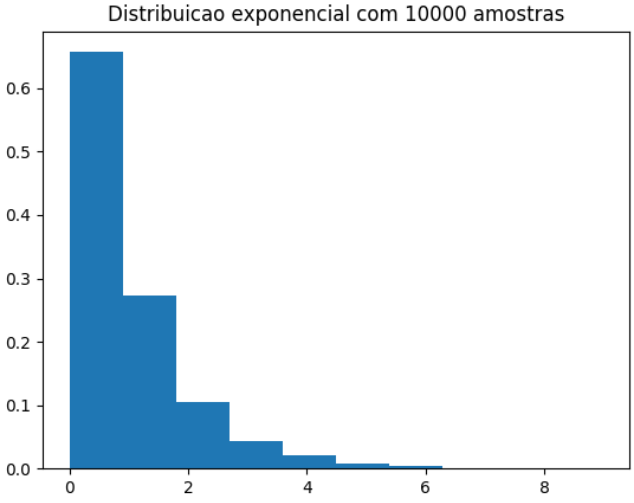

Figura 3.1.10: Histograma da distribuição exponencial com 10000 amostras.

No último teste, com 10000 amostras, fica bem nítido como a variável Y se comporta como uma exponencial.

3.2 Experimento 2

Para os testes do experimento 2 o conjunto de dados original foi dividido em 4 conjuntos, um com todos os 52 estados, um com 28, um com 15 e um com 5. Para apresentar os resultados os gráficos e dados obtidos foram divididos em seções diferentes.

3.2.1 Histogramas

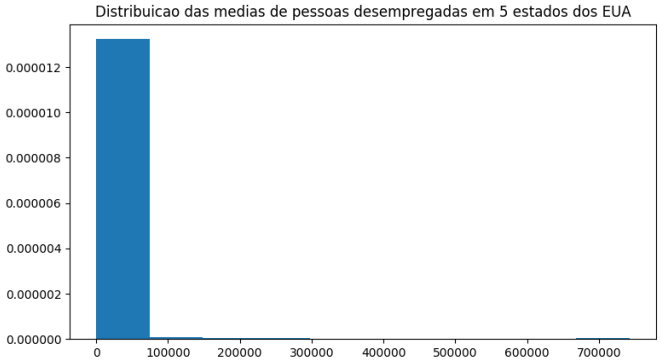

Figura 3.2.1: Histograma dos dados de 5 estados.

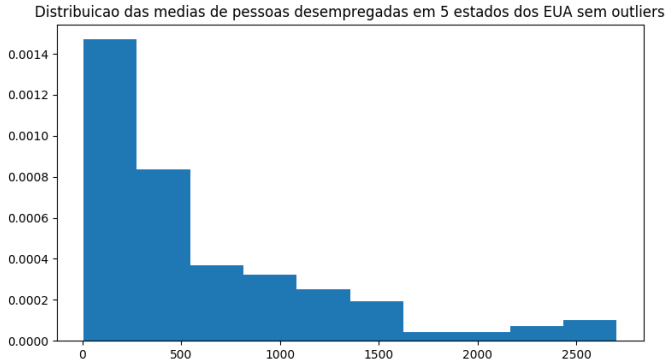

Figura 3.2.2: Histograma dos dados de 5 estados sem outliers.

Figura 3.2.3: Histograma dos dados de 15 estados.

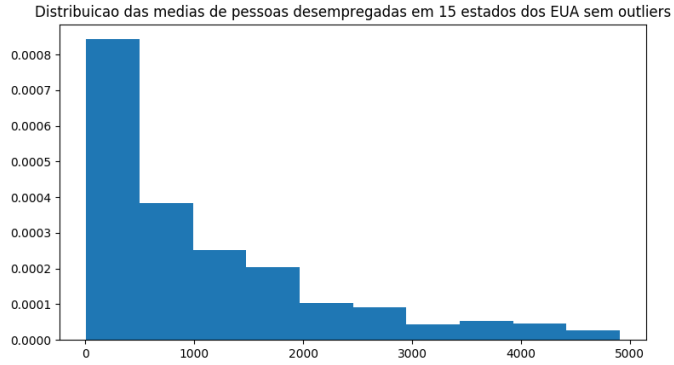

Figura 3.2.4: Histograma dos dados de 15 estados sem outliers.

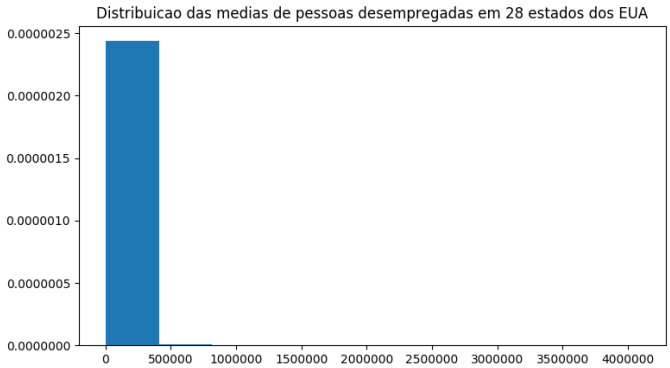

Figura 3.2.5: Histograma dos dados de 28 estados.

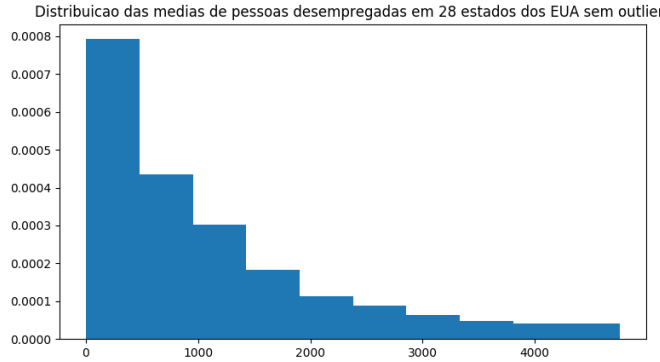

Figura 3.2.6: Histograma dos dados de 28 estados sem outliers.

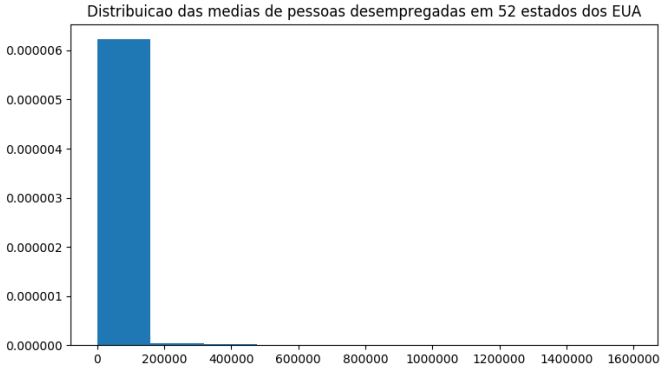

Figura 3.2.7: Histograma dos dados de 52 estados.

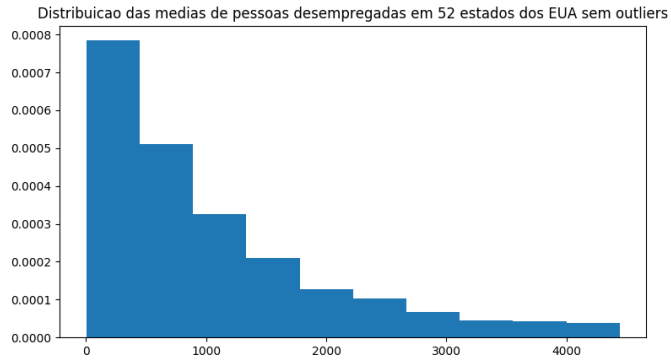

Figura 3.2.8: Histograma dos dados de 52 estados sem outliers.

3.2.2 boxplots

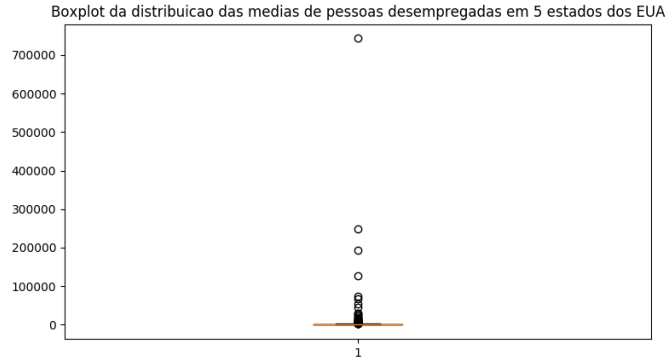

Figura 3.2.9: Boxplot dos dados de 5 estados.

Figura 3.2.10: Boxplot dos dados de 5 estados sem outliers.

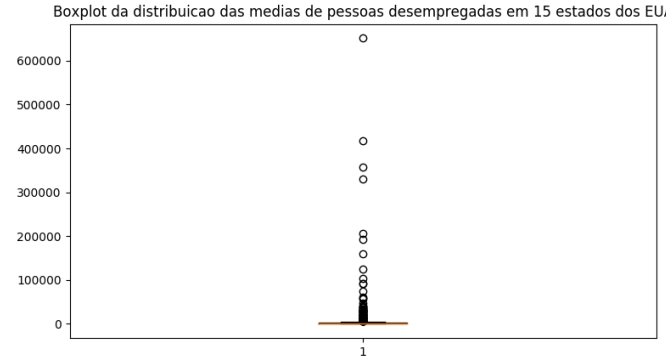

Figura 3.2.11: Boxplot dos dados de 15 estados.

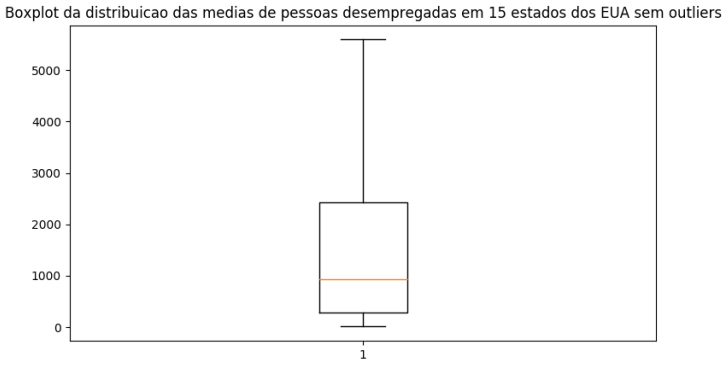

Figura 3.2.12: Boxplot dos dados de 15 estados sem outliers.

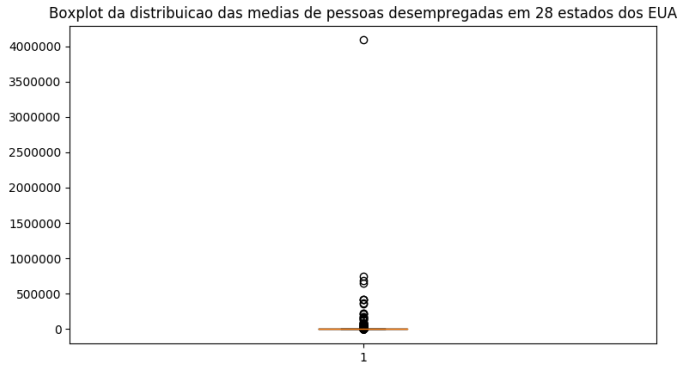

Figura 3.2.13: Boxplot dos dados de 28 estados.

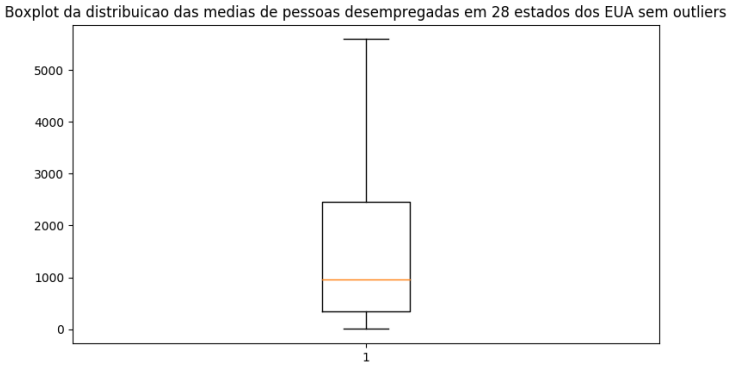

Figura 3.2.14: Boxplot dos dados de 28 estados sem outliers.

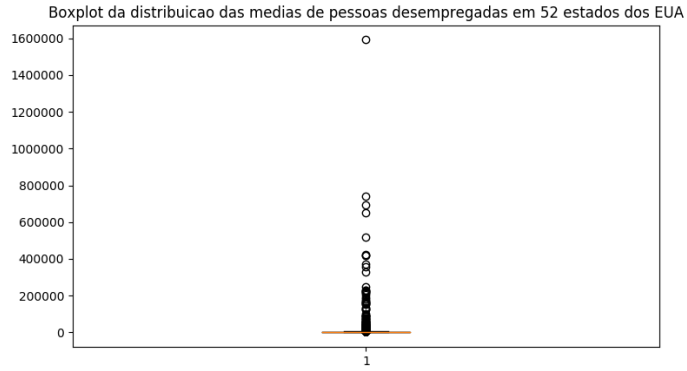

Figura 3.2.15: Boxplot dos dados de 52 estados.

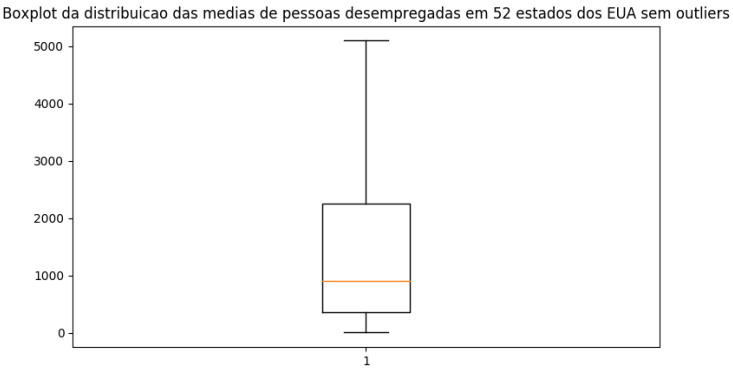

Figura 3.2.16: Boxplot dos dados de 52 estados sem outliers.

3.2.3 Normal probability plot

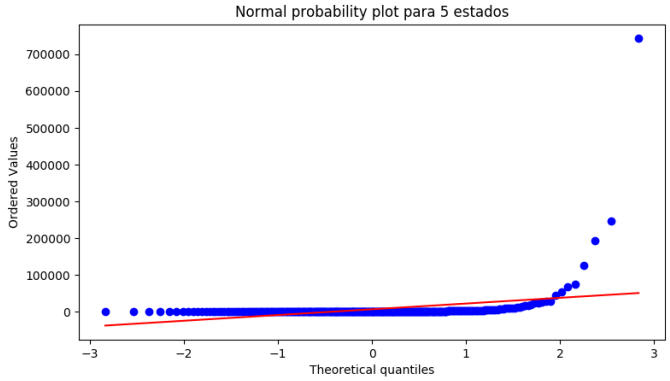

Figura 3.2.17: Normal Probability Plot dos dados de 5 estados.

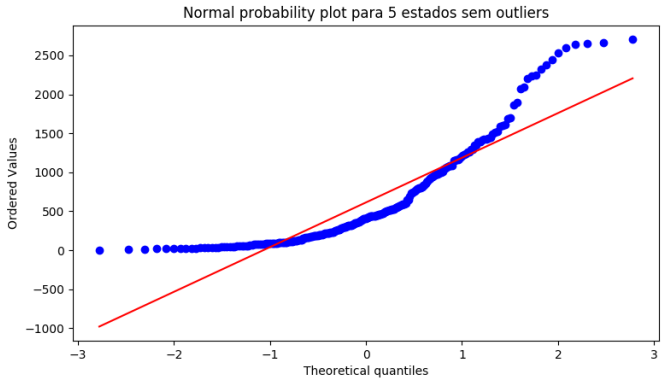

Figura 3.2.18: Normal Probability Plot dos dados de 5 estados sem outliers.

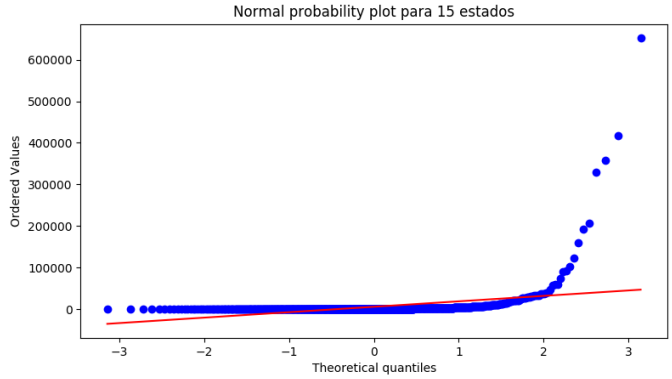

Figura 3.2.19: Normal Probability Plot dos dados de 15 estados.

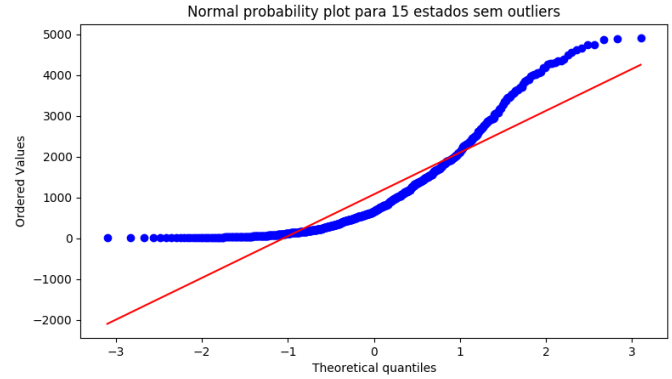

Figura 3.2.20: Normal Probability Plot dos dados de 15 estados sem outliers.

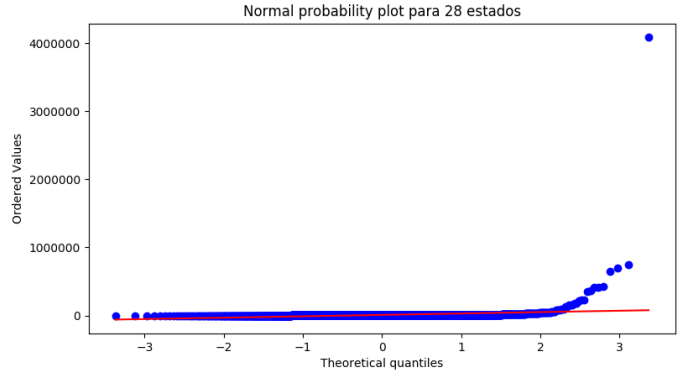

Figura 3.2.21: Normal Probability Plot dos dados de 28 estados.

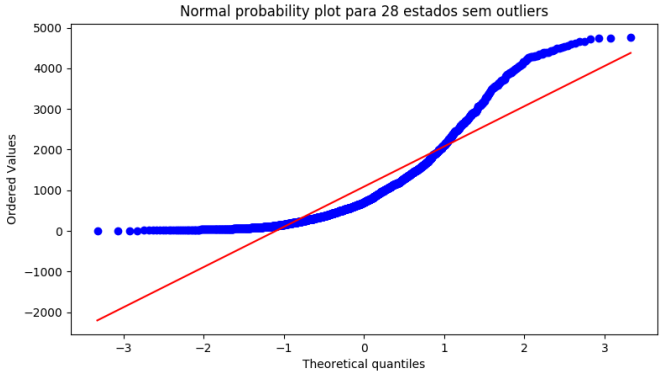

Figura 3.2.22: Normal Probability Plot dos dados de 28 estados sem outliers.

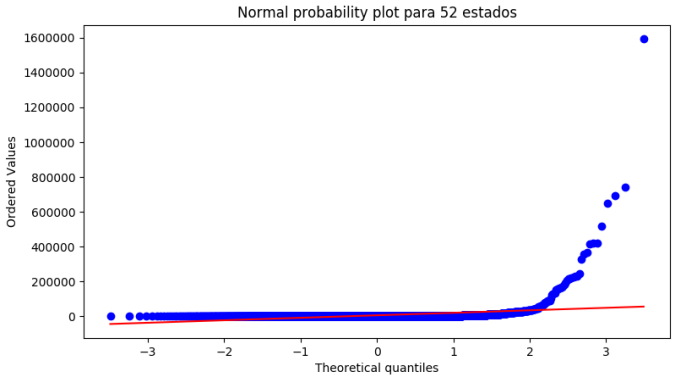

Figura 3.2.23: Normal Probability Plot dos dados de 52 estados.

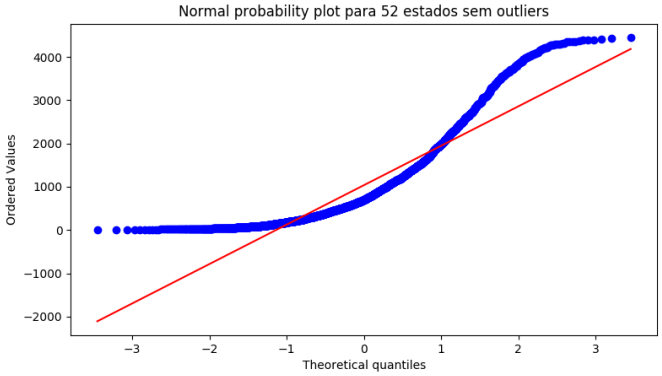

Figura 3.2.24: Normal Probability Plot dos dados de 52 estados sem outliers.

3.2.4 Dados estatísticos

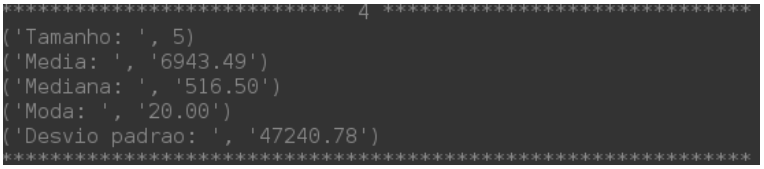

Figura 3.2.25: Dados estatísticos com 5 estados.

Figura 3.2.26: Dados estatísticos com 15 estados.

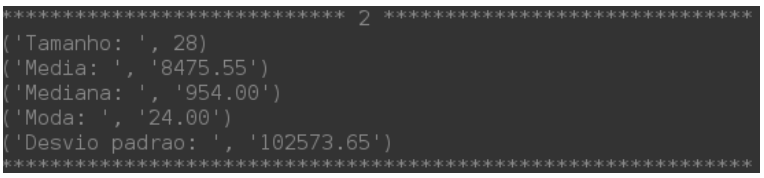

Figura 3.2.27: Dados estatísticos com 28 estados.

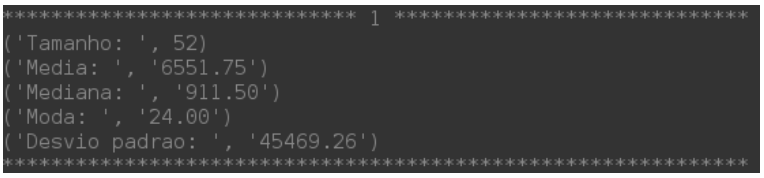

Figura 3.2.28: Dados estatísticos com 52 estados.

4 Conclusão

A partir da análise dos resultados do primeiro experimento conseguimos concluir da boa qualidade do gerador de números pseudo-aleatórios da linguagem Python e que a variável aleatória Y de fato segue uma distribuição exponencial.

Da análise dos resultados do segundo experimento vemos que os dados escolhidos não seguem um padrão normal, e isso é comprovado pelos gráficos de Normal Probability Plot, que ao fazer sem os outliers deixa bem claro i distanciamento do padrão normal, pelos histogramas fica visível como se aproximam mais de um padrão exponencial.