Análise de médias MTBF – Mean Time Between Failures de servidores do Rock in Rio de 2013 a 2023

análise, interpretação dos dados e testes de normalidade.

D. Lapa, M. Nascimento e V. Nobre
Universidade Federal de Pernambuco – Centro de Informática

Keywords: MBTF, servidor, Anderson Darling, Kolmogorov-Smirnov e distribuição normal.

1 Introdução

1.1 Estudo dos dados

O indicador MTBF é usado para estimar a confiabilidade de um sistema ou componente, especialmente onde a confiabilidade é crítica, no caso do estudo foi medido o valor de MTBF de servidores responsáveis pelas vendas de ingressos das edições do Rock in Rio no período de 2013 a 2023.

O MTBF representa a média de tempo que um servidor opera continuamente antes de ocorrer uma falha ou defeito. Geralmente, é expresso em horas e quanto maior o valor do MTBF, maior a confiabilidade do sistema, pois indica que o sistema é menos propenso a falhas. O cálculo do MTBF é relativamente bem simples, envolve o acompanhamento do tempo de operação de um servidor e o registro do número de falhas ou defeitos que ocorrem durante esse período. O cálculo do MTBF consiste em:

$$MTBF = \frac{Tempo\ total\ de\ operação}{N°\ de\ falhas}$$

Com isso, foram coletados durante o período de estudo as MTBF de cada servidor a fim de inferir uma hipótese de que os servidores tiveram médias superior a 3 anos (26280 horas) o que seria o ideal para a permanência dos servidores a fim de evitar custos adicionais com infraestrutura de hardware e em caso de a hipótese nula ser refutada, a organização do evento deveria investir em uma nova infraestrutura de servidores para evitar custos de manutenção mais frequentemente.

2 Metodologia

2.1 Objetivo

O objetivo do estudo é ajudar a organização do evento Rock in Rio a tomar decisão quanto à infraestrutura de servidores que são utilizados para venda de ingressos do evento. Para isso, foi utilizado ferramentas e conceitos estatísticos, além de linguagens de programação para calcular as medidas de centralidade necessárias para inferir e realizar os testes de hipótese tanto para a normalidade dos dados quanto a hipótese do estudo.

2.2 Ferramentas

A utilização da linguagem de programação Python, foi de extrema importância pela facilidade e abrangência de bibliotecas responsáveis pela exibição de gráficos, histogramas, box-plot, etc. Além de realizar os testes de distribuição normal de Anderson Darling e Kolmogorov-Smirnov. Foi utilizado também, o Jupyter Notebook para exibição dos gráficos e a organização do código para melhor visualização.

As bibliotecas utilizadas foram, *Matplotlib* e *Seaborn* para a exibição dos gráficos de QQ-plot, histograma de distribuição de frequência e box plot. Para a realização dos testes de hipótese quanto a normalidade dos dados foi utilizada o modulo *stats* da biblioteca *Scipy* que permitia identificar se os dados seguiam uma distribuição normal.

2.3 Teste de Hipótese Inicial – Distribuição Normal

A normalidade dos dados foi testada de duas maneiras, inicialmente por interpretação de gráficos e posteriormente usando testes de hipótese. O primeiro gráfico gerado a partir dos dados foi o QQ-Plot que envolve a observação de como os pontos se desviam da linha de referência. Se os pontos estiverem próximos à linha de referência, isso sugere que os dados se assemelham à distribuição normal. Se os pontos desviarem significativamente da linha de referência, isso pode indicar que a distribuição teórica não se assemelha a normal.

Logo após, foi realizado a exibição do Histograma de distribuição de frequência com o objetivo de identificar uma curva Gaussiana que seria um indício que a distribuição segue a normalidade. Além do Histograma, foi exibido um gráfico de tipo Box-Plot que separa a distribuição em quartis a partir da mediana, o primeiro quartil corresponde a 25% dos dados, o segundo quartil corresponde a mediana que é equivalente a 50% dos dados e o terceiro quartil corresponde a 75% dos dados. O objetivo da exibição do

gráfico Box-Plot é identificar se o range interquartil (IQR) forma uma simetria entre a mediana e os quartis dando indícios de que a distribuição segue a normalidade.

Entretanto, testes gráficos não tem uma precisão tão grande quanto os testes de hipótese. Portanto, realizar os testes de hipótese foi necessário para obter indícios mais fortes de que a distribuição segue ou não a normalidade rejeitando ou aceitando a hipótese nula. Os testes de hipótese escolhidos foram o de Anderson Darling e o de Kolmogorov-Smirnov, ambos têm a mesmas hipóteses nula e alternativa e são usados para amostras maiores que 50.

$$H_0 = \text{É uma distribuição normal}$$
$$H_a = \text{Não é uma distribuição normal}$$

Os testes foram realizados usando o modulo *stats* da biblioteca *Scipy* que fornecia o algoritmo de cálculo para os testes. O teste de Anderson Darling, parte do pressuposto que não se tem o valor da média e do desvio padrão, assim foi utilizado o método *stats.anderson* que recebe como parâmetros um array contendo os dados e uma string *'norm'* para informar qual tipo de distribuição o teste vai ser realizado, assim retorna o valor da estatística de Anderson Darling, um array contendo os valores críticos, que varia entre 85% a 99% e outro array contendo os níveis de significância referente aos valores críticos.

Tendo esses valores, o teste de hipótese consiste em verificar se a estatística de Anderson Darling é menor do que o nível de significância (alfa) correspondente ao nível de confiança específicado, a hipótese nula de normalidade é aceita, sugerindo que os dados seguem uma distribuição normal. Por outro lado, se a estatística de teste for maior do que o nível de significância (alfa), há evidência suficiente para rejeitar a hipótese nula, indicando que os dados podem não seguir uma distribuição normal.

Já o teste de Kolmogorov-Smirnov, utilizando o método *stats.kstest* recebe como parâmetros um array contendo os dados, o parâmetro *cdf* para indicar qual o tipo de distribuição, parâmetro *args* para informar as medidas de centralidade da distribuição a fim de calcular o p-valor e o valor da estatística de Kolmogorov-Smirnov (*KS_stats*) chamado de estatística D. O método retorna o KS_stats e o p-valor afim de comparar com os mesmos níveis de significância alfa utilizados no teste de Anderson Darling. O algoritmo do teste compara a função de distribuição acumulada (CDF) dos dados observados com a CDF teórica de uma distribuição normal.

Para determinar se a hipótese nula deve ser rejeitada, o teste calcula a estatística D. A estatística D é a maior diferença absoluta entre a CDF empírica dos dados observados

e a CDF teórica da distribuição normal. Quanto maior for a estatística D, maior será a evidência de que os dados não seguem uma distribuição normal.

Para decidir se a hipótese nula deve ser rejeitada, você compara a estatística D com um valor crítico da tabela de Kolmogorov-Smirnov, ajustado para o nível de significância escolhido. Se a estatística D for maior do que o valor crítico correspondente, a hipótese nula de normalidade é rejeitada, indicando que os dados não seguem uma distribuição normal. Se a estatística D for menor do que o valor crítico, não há evidência suficiente para rejeitar a hipótese nula, sugerindo que os dados podem ser aproximadamente normais.

Para encontrar os valores críticos foi calculado por meio da tabela da figura 1 que nos informa a quantidade de amostras e o coeficiente de multiplicação de cada nível de significância (1° linha). Com isso foi criado o método *kolmogorov_smirnov_critico* que calcula o valor crítico de cada nível de significância para amostras maiores que 50, no caso da amostra de estudo foi realizado com tamanho de 1000 amostras, logo após faz a comparação com a estatística D para rejeitar ou anular a hipótese nula.

$n \backslash^{\alpha}$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

Fonte: https://real-statistics.com/statistics-tables/kolmogorov-smirnov-table/
Figura 1. Tabela de cálculo de valores críticos para o teste de Kolmogorov-Smirnov

3 Resultados e Intepretação dos Dados

3.1 Análise Descritiva

Os dados estatísticos coletados são qualitativos contínuos, pois são valores que variam no intervalor de [17528.8, 28721.8]. Foram calculadas as medidas de centralidade e dispersão, a fim de estudar a tendência central da base de dados e usar os valores para cálculos em testes de hipótese. Os valores se encontram na tabela

Usando as medidas de centralidade podemos entender melhor os gráficos gerados a partir da base de dados, como o Histograma de distribuição de frequência (figura 2) e o Box-plot (figura 3) de uma forma mais precisa e assim interpretar os dados de maneira clara e concisa, sem realizar inferências estatísticas ou fazer previsões ainda.

Tabela1. Tabela de medidas estatísticas dos dados

Medição	**Valores**
Média	23132.1581
Mediana	23193.6
Moda	25896.5
Desvio Padrão	3243.94
Variância	10523159.1365
Quartil 25%	20293.775
Quartil 50%	23193.6
Quartil 75%	25870.525
Curtose	-1.1692

3.2 QQ-Plot

O teste QQ (Quantile-Quantile) plot é uma ferramenta gráfica usada na análise estatística para avaliar se a distribuição segue uma distribuição normal. Ele compara os quantis (valores ordenados) dos dados observados com os quantis esperados de uma distribuição normal.

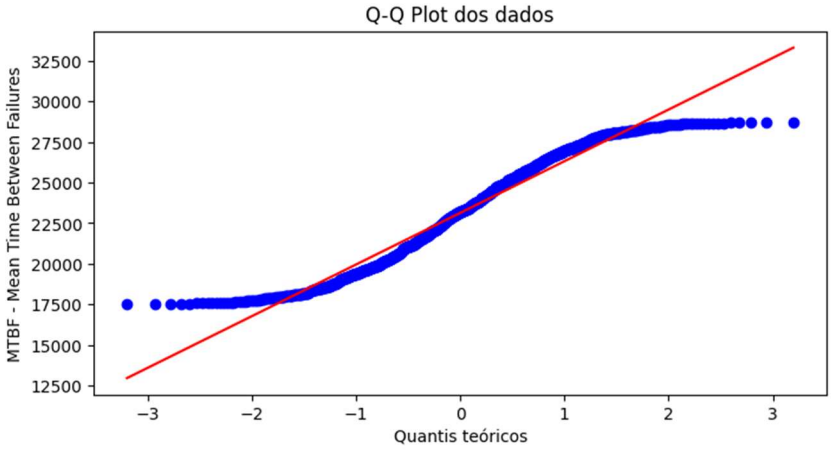

Figura. 2. Gráfico QQ-Plot

A figura 2 compara os quantis teóricos de uma distribuição normal com os quantis observados da distribuição dos dados do estudo. Dessa forma, cada ponto no gráfico corresponde a um par de valores, um do conjunto de dados observados e outro dos quantis teóricos. Se os dados seguirem perfeitamente a distribuição teórica, os pontos formarão uma linha reta de 45° (a linha de referência). Desvios dessa linha indicam desvios da distribuição normal. Portanto, pode-se interpretar que a distribuição do estudo apresenta indícios de que não segue a normalidade por haver consideráveis desvios no gráfico referente a linha de referência do QQ-Plot.

3.3 Histograma de Distribuição de Frequência

O Histograma é um ótimo gráfico para analisar se a distribuição é normal, pois a distribuição normal se assemelha uma distribuição Gaussiana, na qual são bastante semelhantes. No caso do histograma, é possível traçar uma linha referente a continuidade dos dados assim criando uma linha continua em cima das divisões das caixas (bins) do histograma, como é mostrado na figura 3.

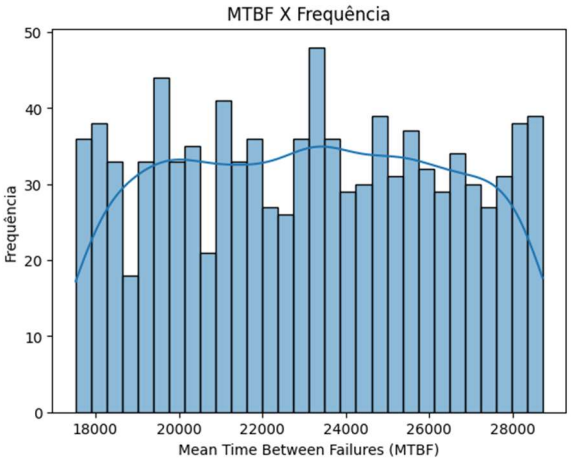

Figura. 3. Histograma de Distribuição de Frequência

Assim, pode-se analisar que tanto a linha contínua não se assemelha a curva Gaussiana quanto a distribuição dos dados não segue uma tendência de Gauss (em forma de pirâmide). Portanto, com a interpretação do gráfico acima, pode-se concluir que há mais um indício que a distribuição não segue a normalidade.

3.4 Box-Plot

O gráfico Box-Plot também é um dos gráficos possíveis de informar algum indício de que a distribuição se assemelhe a uma normal, pela característica da simetria entre os quartis. Se a linha da mediana estiver entre os quartis de maneira que haja uma simetria no range de interquartil (IQR), significa que a distribuição se assemelha a uma distribuição normal.

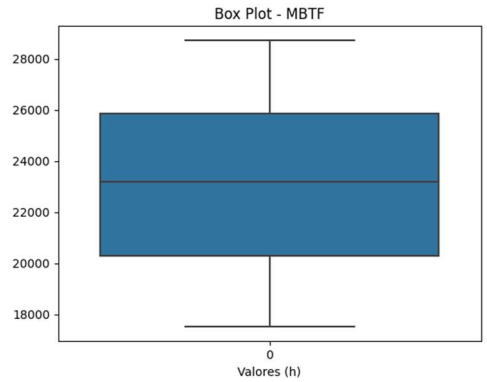

Figura. 4. Box-Plot

Embora, os outros gráficos anteriores tenham informado indícios de que a distribuição não se assemelhe a normal, o gráfico de box-plot dá indícios de que há simetria, portanto, a distribuição pode se assemelhar a normal, se levado em consideração apenas o Box-Plot. Com isso, os testes de gráficos não entregam certezas, nem fortes indícios de normalidade e pela contradição encontrado na interpretação do gráfico Box-plot foi necessário ser feitos testes de hipótese para obter informações mais fortes para o veredito da normalidade dos dados.

3.5 Teste de Hipótese Anderson Darling

Para a realização do teste de Anderson Darling, foi calculado por meio da biblioteca *Scipy* o valor de estatística e comparado com os níveis de significância fornecidos no array do método. O valor da estatística de Anderson Darling encontrado foi 10.088245879678652 e os níveis de significância fornecidos calculados sobre a distribuição foram 0.574, para 85% de confiança, 0.653, para 90% de confiança, 0.784, para 95% de confiança, 0.914, para 97,5% de confiança e 1.088, para 99% de confiança.

Assim, é concluído que temos fortes evidências para rejeitar a hipótese nula visto que o valor de estatística é maior que todos os níveis de significância testados e calculados, portanto, a partir do teste de Anderson Darling temos indícios mais fortes de que a distribuição não segue a normalidade.

```
H0 = É uma distribuição normal
Ha = NÃO é uma distribuição normal
Estatística do teste AD: 10.088245879678652

O resultado da estatistica de Anderson Darling é igual a 10.0882 e o teste:
com 85.0% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Anderson Darling
com 90.0% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Anderson Darling
com 95.0% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Anderson Darling
com 97.5% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Anderson Darling
com 99.0% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Anderson Darling
```

Figura. 5. Terminal do código do teste de Anderson Darling

3.6 Teste de Hipótese Kolmogorov-Smirnov

Para a realização do teste de Kolmogorov-Smirnov, foi calculado por meio da biblioteca *Scipy* o valor de estatística D (*KS_stats*) e com o uso do método *kolmogorov_smirnov_critico* foi encontrado os valores críticos para cada nível de confiança usando a fórmula abaixo para amostras maiores que 50 sendo *k* o coeficiente de cada nível de confiança coletado da tabela da figura 1.

$$ks\_critico = \frac{k}{sqrt(n)}$$

O valor da estatística D encontrado foi 0.06289108423831957 e o p-valor foi de 0.0006998734933681466, já os valores críticos calculados para cada nível de significância foram:

Tabela1. Tabela de valores críticos do teste de Kolmogorov-Smirnov

Nível de Confiança	Nível de Significância	KS Crítico
80%	0,2	0.03392333359945629
85%	0,15	0.03598513863388607
90%	0,1	0.03870153514397071
95%	0,05	0.04294689290274677
99%	0,01	0.05146986365243258

Assim, podemos concluir que para todo KS crítico encontrado o valor da estatística D será maior, além de que o p-valor é suficientemente maior que todos os níveis de significância, portanto, há evidências fortes e suficientes para rejeitar hipótese nula, tendo mais um forte indício de que a distribuição não se assemelha a uma distribuição normal.

```
H0 = É uma distribuição normal
Ha = NÃO é uma distribuição normal
Estatística do teste KS: 0.06289108423831957
Valor P: 0.0006998734933681466

Com 80% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Kolmogorov-Smirnov
Com 85% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Kolmogorov-Smirnov
Com 90% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Kolmogorov-Smirnov
Com 95% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Kolmogorov-Smirnov
Com 99% de confiança, teve evidências para REJEITAR a hipótese nula, segundo o teste de Kolmogorov-Smirnov
```

Figura. 6. Terminal do código do teste de Kolmogorov-Smirnov

4 Conclusão

Contudo, depois de todos os testes gráficos e testes de hipótese realizados e verificado que a distribuição dos dados fornecidos não se assemelha a uma distribuição normal, foi decidido, não seguir adiante com o teste de hipótese que tinha como objetivo analisar se os dados seguiam a média MTBF maior ou igual a 3 anos como hipótese nula, estabelecida pela organização do Rock in Rio, a fim de tomar uma decisão quanto a permanência dos servidores a fim de evitar custos adicionais com infraestrutura e ou investir em uma nova infraestrutura de servidores. A decisão foi tomada, pela quantidade de indícios fortes e visuais de que a distribuição não se assemelhava a normal, assim tornando, incapaz de realizar o teste de hipótese requerido pela organização, pelo motivo da equipe não ter conhecimento de técnicas de teste de hipótese que fujam do escopo das distribuições normais, com base nisso, finalizamos o estudo do caso.

Referências

1. Página de documentação da biblioteca scipy, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html, último acesso em 22/09/2023.
2. Página de documentação da biblioteca scipy, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html, último acesso em 22/09/2023.
3. Blog Real Statitics, Autor: Charles Zaiontz, https://real-statistics.com/statistics-tables/kolmogorov-smirnov-table/, ultimo acesso em 22/09/2023
4. Blog Real Statitics, Autor: Charles Zaiontz, https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/, ultimo acesso em 22/09/2023.
5. Página UEL, https://www.uel.br/projetos/experimental/pages/arquivos/Kolmogorov-Smirnov.html, último acesso em 21/09/2023.
6. Página UEL, https://www.uel.br/projetos/experimental/pages/arquivos/Anderson_Darling.html, último acesso em 21/09/2023.