# Projeto do Curso – Probabilidade e Estatística para Aprendizado de Máquina

**Gabriel Braun**[1]

[1]Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, UFRJ

## Sumário

## 1. Introdução

O objetivo do presente trabalho é comparar as taxas de usos de dados para dois dispositivos de uso comum: Smart-TV e Chromecast. Foram realizadas estatísticas para os dispositivos analisando a taxa de download e upload ao longo do dia, auxiliando provedores de serviço de internet a entender os dados que passam pela rede.

## 2. Estatísticas Gerais

A partir dos dados coletados para os dispositivos Chromecast e Smart-TV. Para as análises posteriores, os campos de taxa de download e taxa de upload foram reescalonados para $\log 10$. A Tabela 1 apresenta a média, variância e desvio padrão para ambos os dispositivos.

**Tabela 1. Média, variância e desvio padrão das taxas de download e upload, sem distinção de horário, para os dispositivos Smart-TV e Chromecast.**

| | Taxa de Upload $[\log(\text{bps}+1)]$ | | Taxa de Dowload $[\log(\text{bps}+1)]$ | |
|---|---|---|---|---|
| | **Smart-TV** | **Chromecast** | **Smart-TV** | **Chromecast** |
| Média | 2,16 | 3,35 | 2,35 | 3,80 |
| Variância | 4,11 | 0,46 | 6,72 | 1,66 |
| Desvio Padrão | 2,03 | 0,68 | 2,59 | 1,29 |

A Tabela 1 mostra que, sem distinção de horário, a taxa de uso de dados dos dispositivos Chromecast é, em média, mais de dez vezes maior que a utilizada pelos dispositivos Smart-TV tanto para download quando para upload. Por outro lado, os dispositivos de Smart-TV apresentam, no geral, variância e desvio padrão muito superior ao dos dispositivos Chromecast.

### 2.1. Histograma

Os histogramas para as taxas de download e upload, sem distinção de horário, para os dispositivos Smart-TV e Chromecast são apresentados na Figura 1. O número de *bins* foi calculado utilizando o método de Sturges:

$$n = 1 + \log_2 N \tag{1}$$

Onde $N$ é o número de dados.

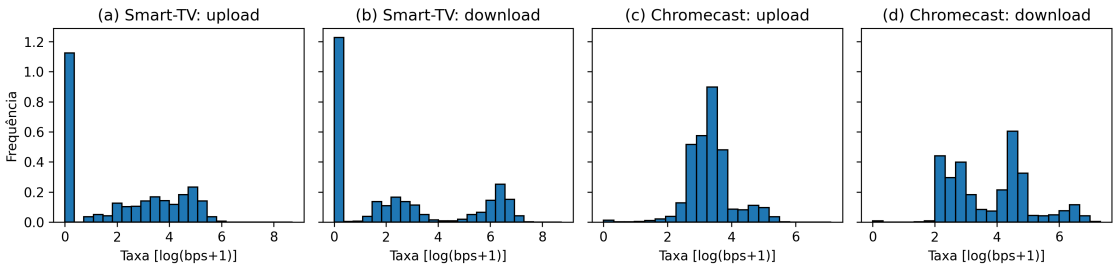


**Figura 1. Histogramas para as taxas de download e upload, sem distinção de horário, para os dispositivos Smart-TV e Chromecast.**

A análise da Figura 1 mostra que os dispositivos do tipo Smart-TV apresentam frequência elevada de dados com taxa baixa ou nula. O mesmo não é observado nos dispositivos do tipo Chromecast, onde a frequência de dados com baixa taxa é aproximadamente nula. Exceto pela grande frequência de dados com taxa baixa nos dispositivos Smart-TV, a distribuição na taxa de download apresenta comportamento aproximadamente unimodal. Já para a taxa de download, a distribuição se aproxima mais de um comportamento bimodal.

Esses dados ajudam a explicar o que foi observado na Tabela 1. A grande frequência de dados com baixa taxa leva a maior variância e desvio padrão observados para os dispositivos Smart-TV.

## 2.2. Função Distribuição Empírica

As funções de distribuição empíricas para as taxas de download e upload, sem distinção de horário para os dispositivos Smart-TV e Chromecast são apresentados na Figura 2.

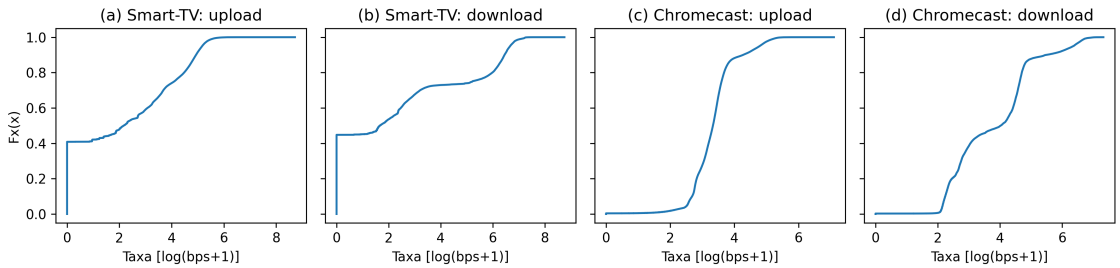


**Figura 2. Função de distribuição empírica para as taxas de download e upload, sem distinção de horário, para os dispositivos Smart-TV e Chromecast.**

As funções de distribuição empírica na Figura 2 revelam o comportamento de característica aproximadamente unimodal para as taxas de upload e bimodal para as taxas de download.

## 2.3. Boxplot

Os boxplots para as taxas de download e upload, sem distinção de horário para os dispositivos Smart-TV e Chromecast são apresentados na Figura 3.

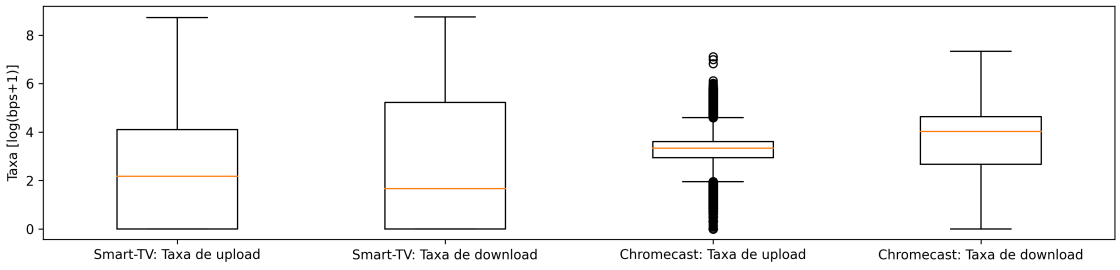


**Figura 3. Boxplots para as taxas de download e upload, sem distinção de horário, para os dispositivos Smart-TV e Chromecast.**

A Figura 3 evidencia o menor desvio padrão e variância das taxas de download e upload dos dispositivos Chromecast.

## 3. Estatísticas por horário

### 3.1. Boxplot

A Figura 4 apresenta os boxplots para as taxas de download e upload, para cada horário ao longo do dia, para os dispositivos Smart-TV e Chromecast.

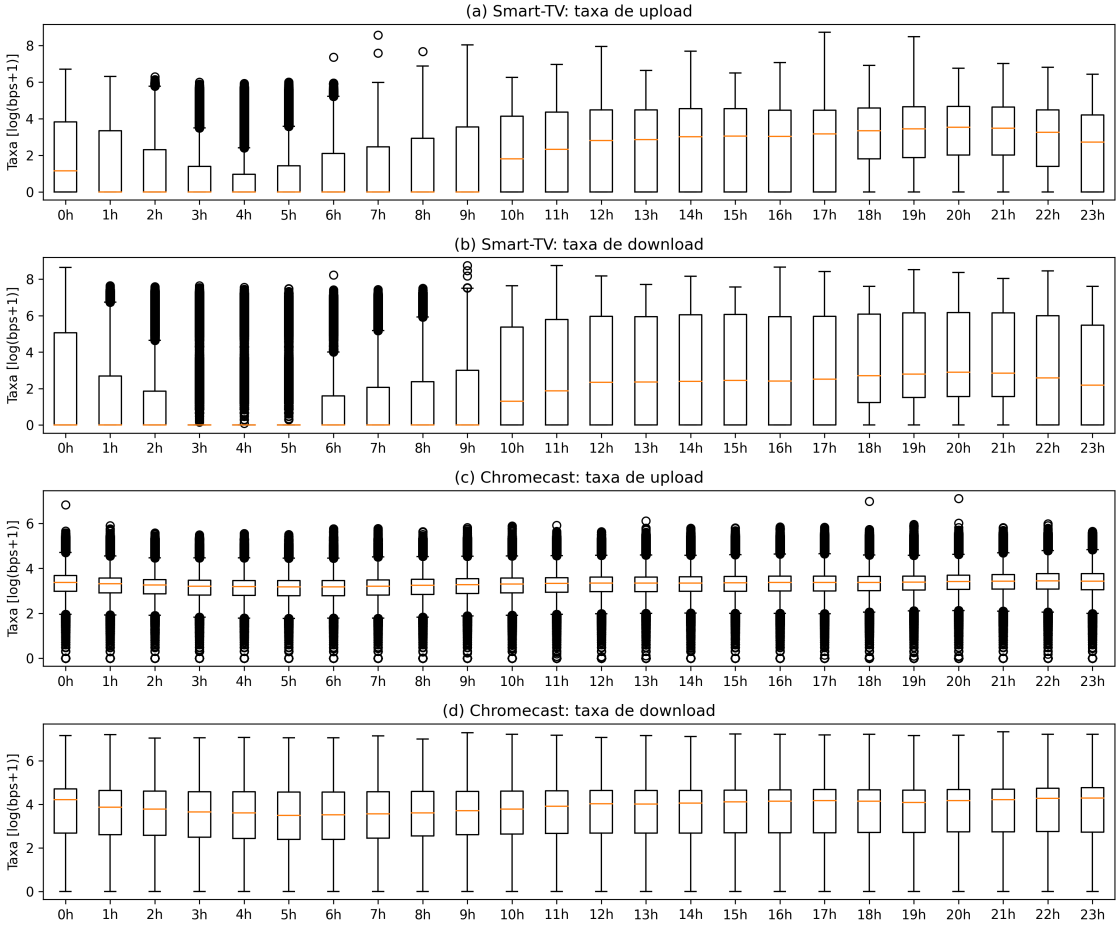


**Figura 4. Boxplots para as taxas de download e upload, para cada horário, para os dispositivos Smart-TV e Chromecast.**

O comportamento geral na Figura 4 é similar ao que é verificado na Figura 3. Os dispositivos Chromecast apresentam pouca variação na distribuição de taxas de download e upload ao longo do dia, diferente dos dispositivos Smart-TV, onde a média é significativamente menor nos horários da madrugada, o que é esperado, já que o número de pessoas utilizando o dispositivo é menor.

Nos horários em que se espera maior uso dos dispositivos, os Chromecasts apresentam menor variância e desvio padrão de uso de dados, cerca de cem vezes menor que a dos dispositivos Smart-TV. Os dispositivos Smart-TV compensam a maior taxa de uso de dados nos horários de pico com a taxa bem menor entre 1h e 8h.

### 3.2. Média, variância e desvio padrão por horário

A Figura 5 apresenta a variação da média, variância e desvio padrão ao longo do dia, para os dispositivos Smart-TV e Chromecast.

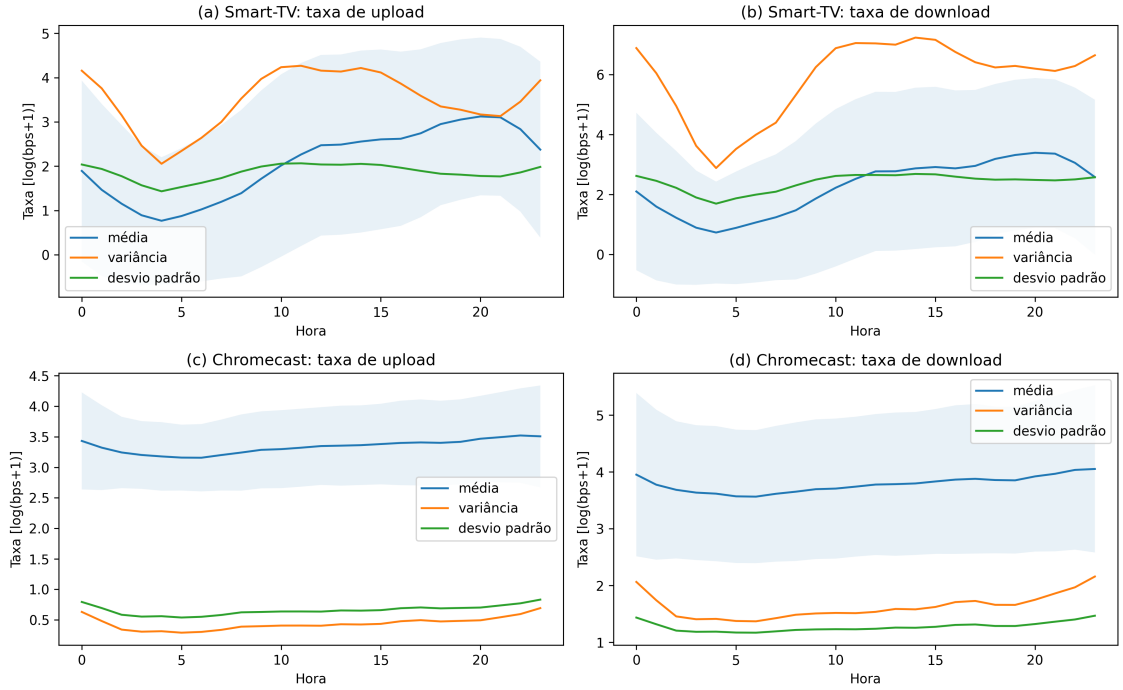


**Figura 5. Média, variância e desvio padrão para as taxas de download e upload ao longo do dia para os dispositivos Smart-TV e Chromecast.**

Na Figura 5 o comportamento evidenciado na Figura 4 pode ser visto claramente. Os dispositivos Smart-TV apresentam grande variância e desvio padrão para a taxa de dados, com essas variando consideravelmente ao longo do dia. Por outro lado, o comportamento dos dispositivos Chromecast é quase constante ao longo do dia.

A região sombreada na figura 5 mostra a faixa de valores para a média mais ou menos um desvio padrão. Podemos esperar uma constância maior para a taxa de download e upload para dispositivos Chromecast nos horários de pico. Para os dispositivos Smart-TV a média é menor ao longo do dia, mas nos horários de pico, o desvio padrão e variância elevados, podem levar ao congestionamento da rede.

No geral, para um mesmo dispositivo, a distribuição da taxa de dowload e de upload ao longo do dia apresenta o mesmo comportamento.

## 4. Caracterizando os horários com maior valor de tráfego

Os horários com maior valor de tráfego foram parametrizados utilizando as distribuições gaussiana (normal) [Dekking et al. 2005],

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{1\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2}$$

e gamma [Dekking et al. 2005],

$$f_{\alpha,\lambda}(x) = \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)} \tag{3}$$

para os dois horários com maior valor da mediana/média de cada taxa coletada para cada tipo de dispositivo. Assim, os dados foram divididos em oito datasets:

- Dataset 1: Horário com a maior mediana da taxa de upload em uma hora, Smart-TV: 20h.
- Dataset 2: Horário com a maior média da taxa de upload em uma hora, Smart-TV: 20h.
- Dataset 3: Horário com a maior mediana da taxa de download em uma hora, Smart-TV: 20h.
- Dataset 4: Horário com a maior média da taxa de download em uma hora, Smart-TV: 20h
- Dataset 5: Horário com a maior mediana da taxa de upload em uma hora, Chromecast: 22h.
- Dataset 6: Horário com a maior média da taxa de upload em uma hora, Chromecast: 22h.
- Dataset 7: Horário com a maior mediana da taxa de download em uma hora, Chromecast: 23h
- Dataset 8: Horário com a maior média da taxa de download em uma hora, Chromecast: 23h

Os horários de maior média e mediada da taxa de download e upload são compatíveis com o horário esperado de pico de utilização desse tipo de dispositivo. Para todos os casos, o horário com a maior média é o mesmo horário com a maior mediana.

Para estimar os parâmetros das distribuições normal e gamma para os oito datasets foi utilizado o método o *maximum likelihood estimator* (MLE). Para a função Gaussiana é possível mostrar [Dekking et al. 2005] que o MLE dos parâmetros para um conjunto de dados $x_1, x_2, \ldots, x_n$ é

$$\hat{\mu} = \overline{x} \quad \text{e} \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4}$$

Onde $\overline{x}$ é a média amostral.

Para a função gamma, a log-likelihood para um conjunto de dados $x_1, x_2, \ldots, x_n$ é dada por:

$$\log p(D|\alpha,\lambda) = (\alpha-1)\sum_{i=1}^{n}\log x_i - n\log\Gamma(\alpha) - n\alpha\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n}x_i \tag{5}$$

$$= n(\alpha-1)\overline{\log x} - n\log\Gamma(\alpha) - n\alpha\log\lambda - \frac{n\overline{x}}{\lambda} \tag{6}$$

O valor de $\lambda$ que maximiza a função é encontrado facilmente:

$$\hat{\lambda} = \frac{\overline{x}}{\alpha} \tag{7}$$

Substituindo o valor de $\hat{\lambda}$ na Equação 6:

$$\log p(D|\alpha, \hat{\lambda}) = n(\alpha - 1)\overline{\log x} - n \log \Gamma(\alpha) - n\alpha \log \overline{x} + n\alpha \log \alpha - n\alpha \qquad (8)$$

Para encontrar o valor $\hat{\alpha}$ que maximiza a log-likelihood, a Equação pode ser maximizada iterativamente pelo método de Newton, convergindo em torno de quatro iterações.

A Figura 6 mostra os histogramas para os oito datasets descritos acima, cada um com as distribuições de probabilidade normal e gamma cujos parâmetros foram calculados pelo método MLE.

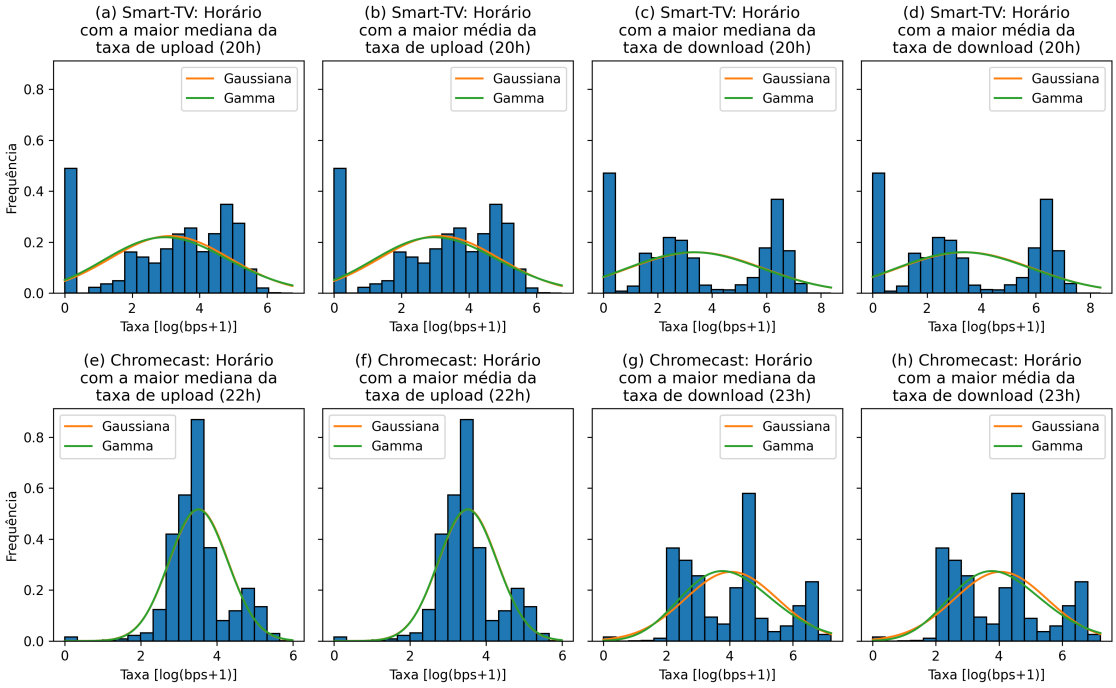


**Figura 6. Histogramas para os oito datasets. Cada um com as distribuições de probabilidade normal e gamma cujos parâmetros foram calculados pelo método MLE.**

Os histogramas da Figura 6 possuem o mesmo comportamento observado nos histogramas da Figura 1. Comparando as Figuras 6 e 1 vemos que a frequência de taxas baixas de dados, característica dos dispositivos Smart-TV, é bem menor nesses horários, o que é esperado para horários com maior utilização dos dispositivos. A distribuição da taxa de upload dos dispositivos Chromecast apresente distribuição com característica trimodal para esse horário, diferente da distribuição da Figura 1, que se aproxima mais de uma bimodal.

## 4.1. Probability Plots

Os gráficos *Probability Plot* podem ser usados para comparar a distribuição real com as distribuições da literatura com os parâmetros obtidos pelo método MLE. Os gráficos *Probability Plot* são apresentados na Figuras 7 e 8.

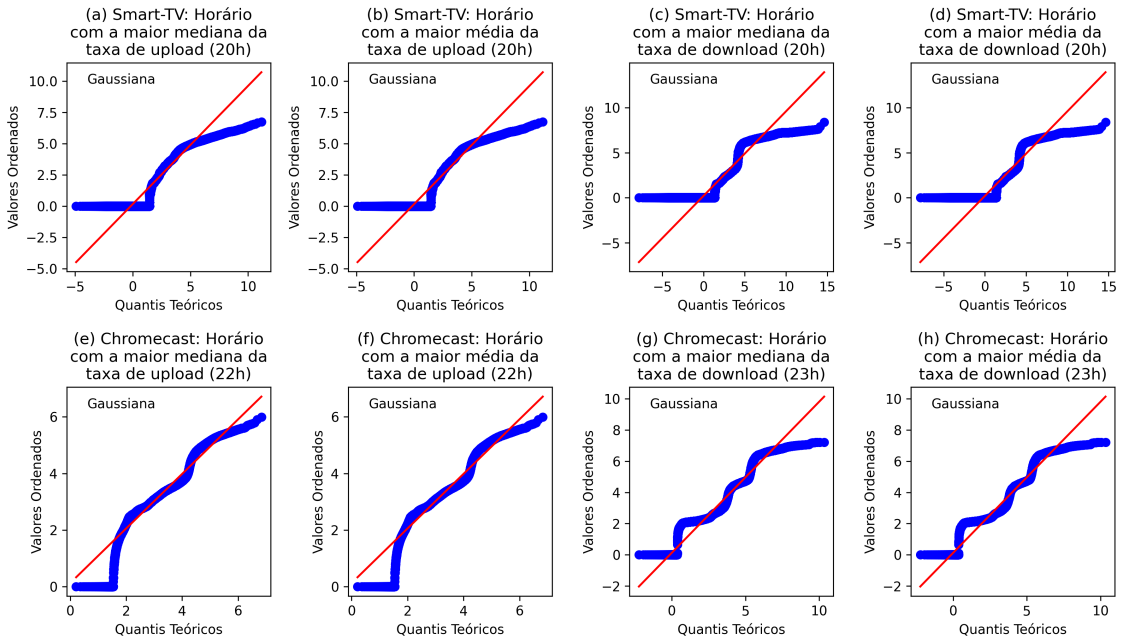


**Figura 7. Probability Plot para os oito datasets comparados com as distribuições de probabilidade normal cujos parâmetros foram calculados pelo método MLE.**

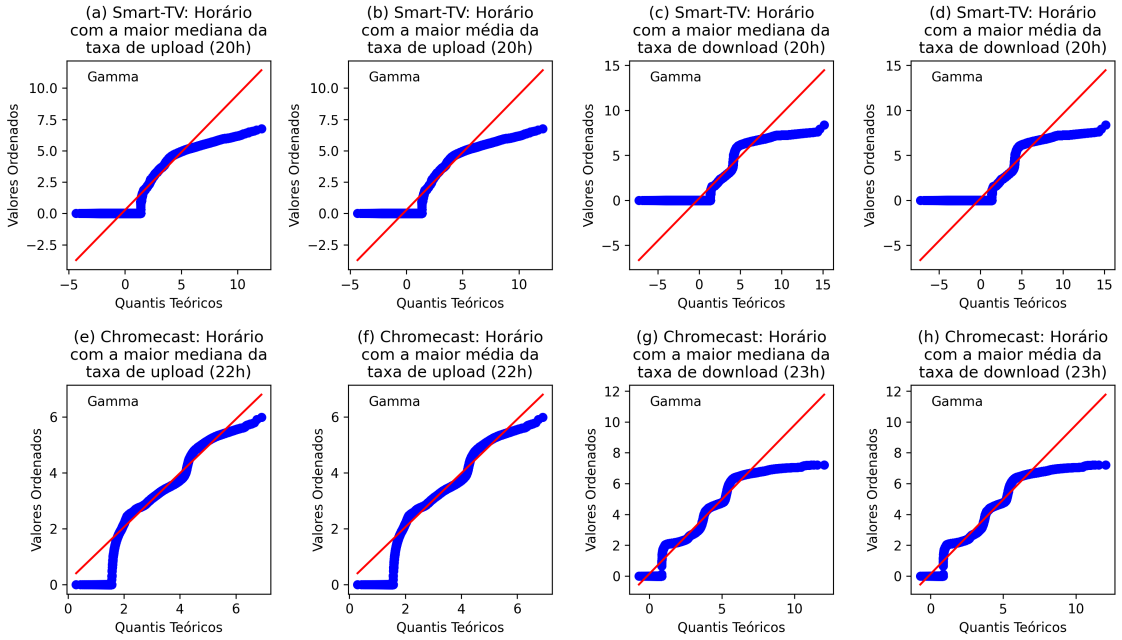


**Figura 8. Probability Plot para os oito datasets comparados com as distribuições de probabilidade gamma cujos parâmetros foram calculados pelo método MLE.**

Para ambos os dispositivos, a distribuição da taxa de download apresenta maior similaridade às distribuições da literatura. Os parâmetros calculados pelo método MLE para as ditribuições normal e gamma levam a distribuições muito similares. Ambas as

distribuições modelam de forma satisfatória a distribuição da taxa de upload dos dispositivos Chromecast. O pico de frequência de taxas muito baixas na distribuição da taxa de upload dos dispositivos Smart-TV leva a distribuições calculadas por MLE com um deslocamento no sentido das taxas mais baixas.

As distribuições para as taxas de download de ambos os dispositivos não apresentam comportamento unimodal e, por isso, não são modeladas satisfatoriamente pelas distribuições normal e gamma.

Comparando as Figuras 7 e 8 vemos que a distribuição gamma se aproxima mais ao comportamento da distribuição da taxa de download dos dispositivos Chromecast para frequências mais baixas, enquanto a distribuição gaussiana é mais próxima do comportamento real para frequências mais altas. Não se nota outras diferenças relevantes entre as distribuições.

## 5. Análise da correlação entre as taxas de upload e download para os horários com o maior valor de tráfego

Os gráficos *scatter plot* e os coeficientes de correlação amostral entre os dados de taxa de download e os dados de taxa de upload para os dois tipos de dispositivo nos horários de maior tráfego são apresentados na Figura 9.

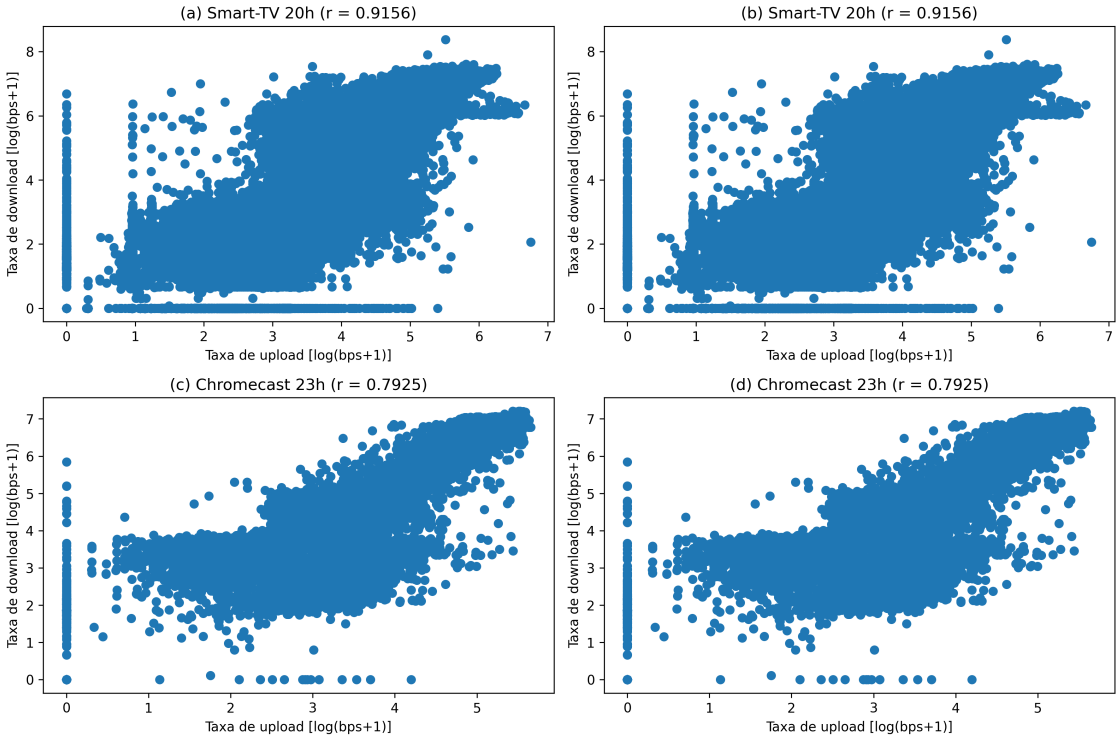


**Figura 9. Gráficos scatter plot e os coeficientes de correlação amostral entre os dados de taxa de download e os dados de taxa de upload para os dois tipos de dispositivo nos horários de maior tráfego são apresentados na Figura 9.**

A análise da Figura 9 mostra que os dispositivos Smart-TV apresentam alta correlação entre as taxas de download e de upload ($r = 0{,}9156$). Os dispositivos Ch-

romecast também apresentam boa correlação entre as taxas de download e de upload ($r = 0{,}7925$), entretanto, essa correlação é menor do que a dos dispositivos Smart-TV.

## 6. Comparação dos dados gerados pelos dispositivos Smart-TV e Chromecast

A estatística G do teste Chi-Square for goodness of fit foi utilizada para avaliar se os dois dispositivos que são usados prioritariamente para assistir vódeo, possuem distribuição de probabilidade das taxas de upload e download semelhante nos horários de maior tráfego. Os resultados obtidos foram:

- Comparação Smart-TV: taxa de upload com maior mediana e Chromecast: taxa de upload com maior mediana. Estatística: 2,9494, *p-value*: 1,0000
- Comparação Smart-TV: taxa de upload com maior média e Chromecast: taxa de upload com maior média. Estatística: 2,9494, *p-value*: 1,0000
- Comparação Smart-TV: taxa de download com maior mediana e Chromecast: taxa de download com maior mediana. Estatística: 3,5924, *p-value*: 1,0000
- Comparação Smart-TV: taxa de download com maior média e Chromecast: taxa de download com maior média. Estatística: 3,5924, *p-value*: 1,0000

A partir desses resultados podemos concluir que a distribuição de probabilidade na taxa de download e na taxa de upload é semelhante nos horários de maior tráfego. O que era esperado, dada a correlação entre essas taxas evidenciada na Figura 9.

## 7. Conclusão

A partir das estatísticas analisadas podemos entender as principais diferenças entre os dispositivos Smart-TV e Chromecast. Os dispositivos Smart-TV apresentam grande variabilidade na média da taxa de dados ao longo do dia. Além disso, apresentam grande variância e desvio padrão num mesmo horário. A média da taxa de dados é consideravelmente maior nos horários de pico no uso desse tipo de dispositivo e menor durante a madrugada. Esse tipo de comportamento leva a suposição de que esse dispositivo faz as requisições dos dados assim que esses são requisitados pelo usuário.

Por outro lado, os dispositivos Chromecast apresentam uma distribuição da taxa de dados que varia pouco ao longo do dia. Num mesmo horário, os dados também apresentam menor variância e desvio padrão. Provavelmente, os dispositivos Chromecast utilizam algorítimos de inteligência artificial para identificar quais conteúdos são prováveis de serem requisitados pelo usuário, e usa os horários de menor uso para requisitar previamente esses dados, mantendo a taxa de download e upload constante ao longo do dia. Essa estratégia permite que os dispositivos Chromecast atinjam uma taxa de dados em horário de pico com menor variabilidade. Um usuário do Chromecast deve ter menos experiências de lentidão nos horários de pica do que um usuário das Smart-TV. Em contrapartida, os dispositivos Chromecast utilizam em média uma taxa de dados superior aos dispositivos Smart-TV.

O repositório `https://github.com/gpbraun/projeto-probest` contem o código utilizado no trabalho.

## Referências

Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P., and Meester, L. E. (2005). *A modern introduction to probability and statistics*. Springer Texts in Statistics. Springer, London, England, 1 edition.