



UNIVERSIDADE FEDERAL DO ABC  
CENTRO DE MATEMÁTICA, COMPUTAÇÃO E COGNIÇÃO  
BACHARELADO EM CIÊNCIAS E HUMANIDADES

Caio César Carvalho Ortega, Carolina Horta Cattaneo

*Trabalho de Introdução à Probabilidade e Estatística*

São Bernardo do Campo, SP  
2019

Caio César Carvalho Ortega, Carolina Horta Cattaneo

*Trabalho de Introdução à Probabilidade e Estatística*

Trabalho do Bacharelado em Ciências e Humanidades da UFABC entregue como parte da disciplina de Introdução à Probabilidade e Estatística.

Orientador: Prof. Dr. Antonio Sergio Munhoz

São Bernardo do Campo, SP  
2019

# Resumo

Este trabalho analisa 100 amostras aleatórias de avaliações para a empresa MercadoLivre na plataforma InfoJobs, aplicando os conhecimentos adquiridos ao longo do curso, tais como elaboração de histogramas, construção de tabelas de frequências, distribuições, entre outros.

**Palavras-chaves:** estatística, probabilidade, histograma, dados, análise.

# Sumário

	<b>Sumário</b>	<b>3</b>
	<b>Lista de ilustrações</b>	<b>4</b>
	<b>Lista de tabelas</b>	<b>5</b>
<b>1</b>	<b>INTRODUÇÃO</b>	<b>6</b>
<b>2</b>	<b>CONSIDERAÇÕES METODOLÓGICAS</b>	<b>7</b>
<b>2.1</b>	<b>Intervalo de confiança</b>	<b>7</b>
<b>2.2</b>	<b>Captura e organização dos dados</b>	<b>8</b>
2.2.1	Dados do InfoJobs	9
2.2.2	Dados do UFABC Next	10
<b>3</b>	<b>ANÁLISE DOS DADOS</b>	<b>12</b>
<b>3.1</b>	<b>Dados do InfoJobs</b>	<b>12</b>
<b>3.2</b>	<b>Dados do UFABC Next</b>	<b>13</b>
<b>4</b>	<b>CONCLUSÃO</b>	<b>21</b>
	<b>REFERÊNCIAS</b>	<b>22</b>
	<b>Glossário</b>	<b>23</b>

# Lista de ilustrações

Figura 1 – Histograma para a variável “Frequência estrelas geral” . . . . .	13
Figura 2 – Histograma para a variável “Ambiente de trabalho” . . . . .	14
Figura 3 – Histograma para a variável “Benefícios” . . . . .	15
Figura 4 – Histograma para a variável “Conciliação vida familiar” . . . . .	16
Figura 5 – Histograma para a variável “Oportunidade de promoção” . . . . .	17
Figura 6 – Histograma para a variável “Aprova a diretoria” . . . . .	18
Figura 7 – Histograma para a variável “Funcionário” . . . . .	19
Figura 8 – Histograma para a variável “Recomenda a um amigo” . . . . .	20

# Lista de tabelas

Tabela 1 – Tipos de variáveis para o MercadoLivre . . . . .	9
Tabela 2 – Tipos de variáveis para o MercadoLivre . . . . .	11
Tabela 3 – Tabelas de frequências para as variáveis discretas (amostras do InfoJobs)	12
Tabela 4 – Tabelas de frequências para as variáveis binomiais (amostras do InfoJobs)	12

# 1 Introdução

Este trabalho se propõe a aplicar os conhecimentos discutidos ao longo do curso, expandindo uma das primeiras atividades aplicadas, que consistiu na coleta, tratamento e análise de dados do site InfoJobs<sup>1</sup> para a empresa MercadoLivre.

A partir de uma amostra aleatória contendo 100 mensagens deixadas no site (ou seja, 100 amostras), estas foram disponibilizadas *online* por meio do *software* Google Sheets, além disso, foram elaborados gráficos, tais como um histograma.

Finalmente, a estratégia adotada para o MercadoLivre foi adaptada para a UFABC, utilizando dados da plataforma UFABC Next<sup>2</sup>, que suplantou o antigo UFABC Help!<sup>3</sup>, parametrizando valores com o intuito de proporcionar a melhoria dos planos de ensinos das disciplinas observadas.

---

<sup>1</sup> <<https://www.infojobs.com.br/mercado-livre/vagas>>

<sup>2</sup> <<https://ufabcnext.com>>

<sup>3</sup> <<https://www.ufabchelp.me/>>

## 2 Considerações metodológicas

### 2.1 Intervalo de confiança

O primeiro aspecto metodológico a merecer destaque é a adoção da Regra dos 100, que nada mais é do que o estabelecimento de um intervalo de confiança para 100 amostras. Considerando uma população de 200, seria possível atingir um intervalo de confiança de 93,43% com erro de 11,24%. A exemplo da Regra dos 50 fornecida nas especificações do trabalho<sup>1</sup>, foi elaborada planilha análoga para a Regra dos 100 utilizando o *software* LibreOffice.org Calc, com a posterior carga e conversão no disco virtual do Google (Google Drive), permitindo acesso remoto por meio do *software* Google Sheets, por meio do endereço <[https://docs.google.com/spreadsheets/d/1EVE47HOMK7rcjOnfLdPdV3hR6\\_NcSbeoKX3M4Z51jJk/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1EVE47HOMK7rcjOnfLdPdV3hR6_NcSbeoKX3M4Z51jJk/edit?usp=sharing)>. A Regra dos 100 foi baseada na lógica da Regra dos 50 fornecida no escopo do trabalho disponibilizado eletronicamente na plataforma de EAD Tidia da UFABC por meio do endereço <[https://docs.google.com/spreadsheets/d/1o\\_8Owjhj2WIb7zI8xsssiL\\_a5Xm\\_HvKiqt8Aqjy3Hg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1o_8Owjhj2WIb7zI8xsssiL_a5Xm_HvKiqt8Aqjy3Hg/edit?usp=sharing)>.

As planilhas estão construídas adotando o conceito de distribuição hipergeométrica, adequada para extrações casuais de uma população que são feitas sem reposição, dividida em dois atributos. O termo *split* (divisão, traduzindo livremente do inglês para o português brasileiro) está presente nas planilhas em virtude dessa divisão. A distribuição hipergeométrica é conceituada por Morettin e Bussab (2017, p. 147) como:

Essa distribuição é adequada quando consideramos extrações casuais feitas sem reposição de uma população dividida segundo dois atributos. Para ilustrar, considere uma população de  $N$  objetos,  $r$  dos quais têm o atributo  $A$  e  $N - r$  têm o atributo  $B$ . Um grupo de  $n$  elementos é escolhido ao acaso, sem reposição. Estamos interessados em calcular a probabilidade de que esse grupo contenha  $k$  elementos com o atributo  $A$ .

O cálculo da probabilidade mencionada por Morettin e Bussab (2017) é dado pela equação a seguir, utilizando o princípio multiplicativo:

$$p_k = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}} \quad (2.1)$$

<sup>1</sup> <[https://docs.google.com/spreadsheets/d/1o\\_8Owjhj2WIb7zI8xsssiL\\_a5Xm\\_HvKiqt8Aqjy3Hg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1o_8Owjhj2WIb7zI8xsssiL_a5Xm_HvKiqt8Aqjy3Hg/edit?usp=sharing)>



Finalmente, conforme Morettin e Bussab (2017, p. 147), “os pares  $(k, p_k)$  constituem a distribuição hipergeométrica de probabilidades”, sendo a variável aleatória  $X$  “o número de elementos na amostra que têm o atributo  $A$ , então  $P(X = k) = p_k$ ”.

Como aponta Field (2009, p. 45), o valor do intervalo de confiança se traduz na ideia de que ele representa quantos intervalos possuem o valor real da média da população, em outras palavras, o valor indica a confiabilidade da estimativa:

Tipicamente, se prestarmos atenção aos intervalos de confiança de 95% e, algumas vezes, aos intervalos de confiança de 99%, veremos que eles têm interpretações semelhantes: são limites construídos para que em certa percentagem das vezes (seja 95% ou 99%) o valor real da média da população esteja dentro desses limites. Assim, quando você tiver um intervalo de confiança de 95% para uma média, pense nele assim: se selecionarmos 100 amostras, calcularmos a média e, depois, determinarmos o intervalo de confiança para aquela média (...), 95% dos intervalos de confiança conterão o valor real da média da população.

Na abordagem do tema feita por Field (2009, p. 45-47), o cálculo do intervalo de confiança está intimamente ligado com os escores- $z$ , porque um intervalo de confiança de 95%, por exemplo, corresponde ao valor de  $z$  de 1,96, colocando os limites do intervalo de confiança em  $-1,96$  e  $+1,96$ , correspondendo, portanto, a uma distribuição normal com média 0 e desvio padrão 1 (FIELD, 2009, p. 45). Sabendo-se que a conversão de escores em escores- $z$  é feita por meio da Equação 2.2, sendo  $z$  o escore- $z$ ,  $X$  o valor,  $\bar{X}$  a média e  $s$  o desvio padrão (FIELD, 2009, p. 41).

$$z = \frac{X - \bar{X}}{s} \quad (2.2)$$

Sabendo que a média está sempre no intervalo de confiança, o cálculo deste é feito reorganizando as equações, mantendo a mesma lógica empregada anteriormente (FIELD, 2009, p. 47).

$$z \times s = X - \bar{X} \quad (2.3)$$

$$z \times s + \bar{X} = X \quad (2.4)$$

## 2.2 Captura e organização dos dados

A coleta/captura dos dados ocorreu sem o emprego de técnicas de programação ou de automação para fazer o *scrapping* dos dados<sup>2</sup>. Os passos adotados assim podem ser sumarizados:

<sup>2</sup> A terminologia pode ser traduzida livremente para o português brasileiro como “raspagem” e é empregada em outros artigos, como em Junior (2012, p. 212)

1. Abertura do endereço desejado da WWW no navegador;
2. Identificação das variáveis desejadas;
3. Estruturação de uma planilha capaz de acomodar os dados das variáveis verticalmente, sendo uma variável por coluna, um valor célula, de maneira que para a variável  $V_1$  na coluna  $A$ , possuindo esta três valores, estes serão acomodados nas células  $A2$ ,  $A3$  e  $A4$ , ficando a célula  $A1$  reservada para o título ou nome da variável;
4. Preenchimento da planilha, transferindo os dados por mecanismos de cópia e colagem nativos do sistema operacional.

A classificação das variáveis adotou a conceituação realizada por Pinheiro et al. (2009, p. 6):

**Variável qualitativa nominal** ou **categórica** — seus valores possíveis são diferentes categorias não-ordenadas, em que cada observação pode ser classificada. Exemplos: raça, nacionalidade, área de atividade.

**Variável qualitativa ordinal** — seus valores possíveis são diferentes categorias ordenadas, em que cada observação pode ser classificada. Exemplos: classe social, nível de instrução.

**Variável quantitativa discreta** — seus valores possíveis são em geral resultados de um processo de contagem. Exemplos: número de filhos, número de séries escolares cursadas com aprovação.

**Variável quantitativa contínua** — seus valores podem ser expressos através de números reais e varrem uma escala contínua de medição. Exemplos: renda mensal, peso, altura.

### 2.2.1 Dados do InfoJobs

A Tabela 2 classifica as variáveis para a amostra de 100 avaliações do site InfoJobs para a empresa MercadoLivre. Foram coletadas as avaliações feitas pelos usuários do site no período entre 18 de Fevereiro de 2019 e 30 de Maio de 2019.

A fim de que a amostra tivesse resultados mais homogêneos e relevantes, os dados foram previamente filtrados no próprio site pela localidade, fazendo com que fossem apresentados apenas avaliações do estado de São Paulo.

Tabela 1 – Tipos de variáveis para o MercadoLivre

Nome da variável	Tipo de variável	Subtipo de variável
Estrelas	Quantitativa	Discreta
Global	Quantitativa	Discreta
Oportunidade de promoção	Quantitativa	Discreta
Ambiente de trabalho	Quantitativa	Discreta
Conciliação com a vida familiar	Quantitativa	Discreta

Continua na próxima página

**Tabela 1 – continuado da página anterior**

<b>Nome da variável</b>	<b>Tipo de variável</b>	<b>Subtipo de variável</b>
Benefícios	Quantitativa	Discreta
Recomenda a Empresa a um amigo	Qualitativa	Nominal
Aprova a Diretoria	Qualitativa	Nominal
Funcionário/Ex sim/não	Qualitativa	Nominal
Data	Qualitativa	Ordinal
Cargo	Qualitativa	Nominal
Comentário	Qualitativa	Nominal
Prós	Qualitativa	Nominal
Contras	Qualitativa	Nominal
Dica a Diretoria	Qualitativa	Nominal

Elaboração própria

### 2.2.2 Dados do UFABC Next

No caso do MercadoLivre, a obtenção das 100 amostras não significou grandes entraves, por outro lado, a obtenção de avaliações de alunos do sistema UFABC Next exigiu a observação de mais de um docente, mais precisamente, de 19 (dezenove) professores:

- A. M. Timpanaro;
- A. M. Veneziani;
- A. Magalhães;
- A. P. de Oliveira Junior;
- A. S. Munhoz;
- B. Marin;
- C. S. dos Santos;
- E. Alejandra;
- Ignat F.;
- L. A. da Silva;
- P. J. P. Martinez;
- P. M E. Claessens;

- R. H. A. H Jacobs;
- R. M. Coutinho;
- R. Venegeroles;
- S. Camargo;
- T. L. Ritchie;
- V. Marvulle;
- V. Perchine.

Tabela 2 – Tipos de variáveis para o MercadoLivre

Nome da variável	Tipo de variável	Subtipo de variável
Professor	Qualitativa	Nominal
Cobra Presença	Quantitativa	Discreta
Conceito	Qualitativa	Ordinal
Comentário P/N	Qualitativa	Nominal
Didática P/N	Qualitativa	Nominal

Elaboração própria

### 3 Análise dos dados

Como veremos a seguir, foram realizados três análises: (i) teste de aleatoriedade com determinado número de parâmetros para cada amostra; (ii) probabilidade de obtenção de determinado valor para determinada variável; e (iii) elaboração de tabelas de frequência e histogramas em relação aos parâmetros, observando o formato da curva; seleção de outro parâmetro binominal para verificar sua significância.

Os análises foram realizadas para as duas amostras, uma vez que este trabalho optou por ir além dos dados do InfoJobs.

#### 3.1 Dados do InfoJobs

Como descrito na Capítulo 2, a análise dos dados envolveu a elaboração de tabelas de frequências consoante as variáveis selecionadas e a construção de histogramas, com vistas à análise do comportamento da curva.

A elaboração das tabelas de frequências para as variáveis qualitativas discretas e binomiais pode ser conferida por meio do endereço <[https://docs.google.com/spreadsheets/d/16I1IYcc7LitdmlR3Kj3LKNXpJd9\\_Y8hIAKpe8Y8hAHM/edit#gid=1166636790](https://docs.google.com/spreadsheets/d/16I1IYcc7LitdmlR3Kj3LKNXpJd9_Y8hIAKpe8Y8hAHM/edit#gid=1166636790)> ou a seguir, por meio da Tabela 3 e da Tabela 4:

Tabela 3 – Tabelas de frequências para as variáveis discretas (amostras do InfoJobs)

Quantidade de Estrelas	Global		Oportunidade de promoção		Ambiente de trabalho		Conciliação com a vida familiar		Benefícios	
	Frequência	Porcentagem	Frequência	Porcentagem	Frequência	Porcentagem	Frequência	Porcentagem	Frequência	Porcentagem
1	1	1%	4	4%	1	1%	4	4,00%	2	2%
2	1	1%	5	5%	2	2%	0	0,00%	1	1%
3	6	6%	29	29%	6	6%	13	13,00%	5	5%
4	19	19%	23	23%	21	21%	23	23,00%	9	9%
5	73	73%	39	39%	70	70%	54	54,00%	82	82%
Total geral	100	100%	100	100%	100	100%	100	100,00%	100	100%

Elaboração própria

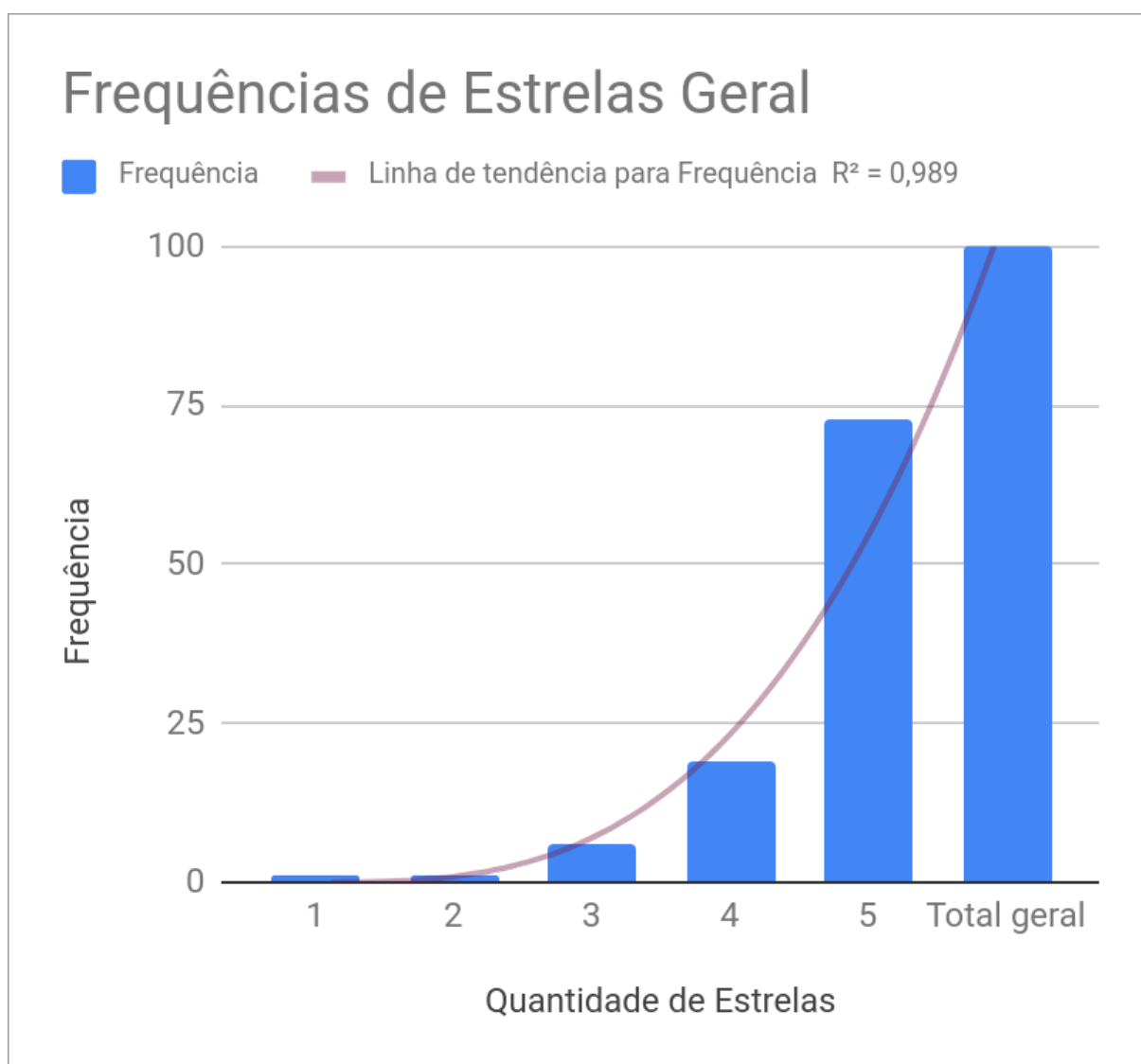
Tabela 4 – Tabelas de frequências para as variáveis binomiais (amostras do InfoJobs)

Sim/Não	Recomenda a Empresa a um amigo		Aprova a Diretoria		Situação	É funcionário ou Ex-Funcionário?	
	Frequência	Porcentagem	Frequência	Porcentagem		Frequência	Porcentagem
Sim	97	97%	80	80%	Funcionário	41	41%
Não	3	3%	19	19%	Ex-funcionário	59	59%
Total geral	100	100%	100	100%	Total geral	100	100%

Elaboração própria

Os histogramas para variáveis podem ser conferidos nas figuras a seguir.

Figura 1 – Histograma para a variável “Frequência estrelas geral”



Elaboração própria

### 3.2 Dados do UFABC Next

Figura 2 – Histograma para a variável “Ambiente de trabalho”



Elaboração própria

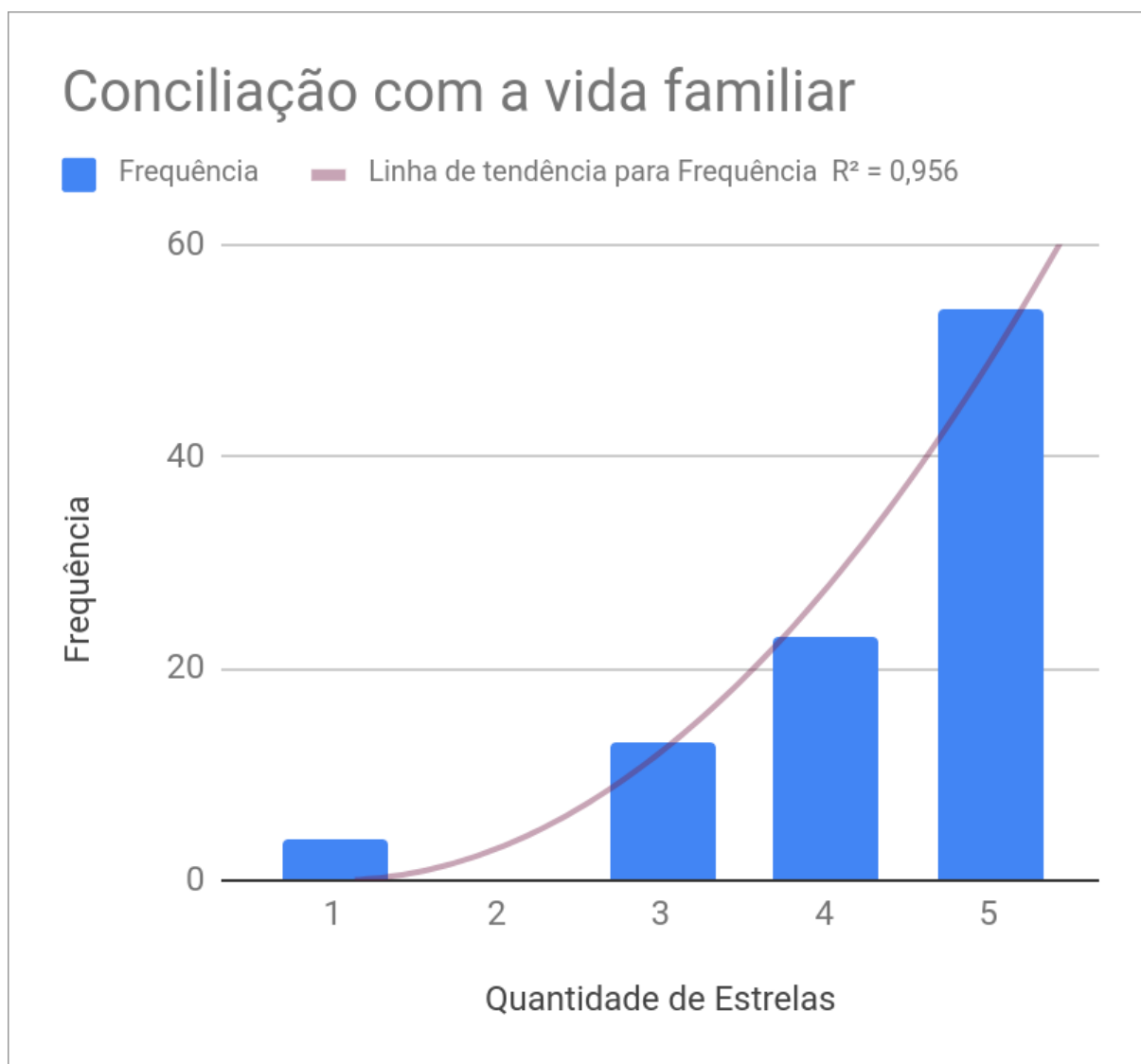
Figura 3 – Histograma para a variável “Benefícios”



Elaboração própria



Figura 4 – Histograma para a variável “Conciliação vida familiar”



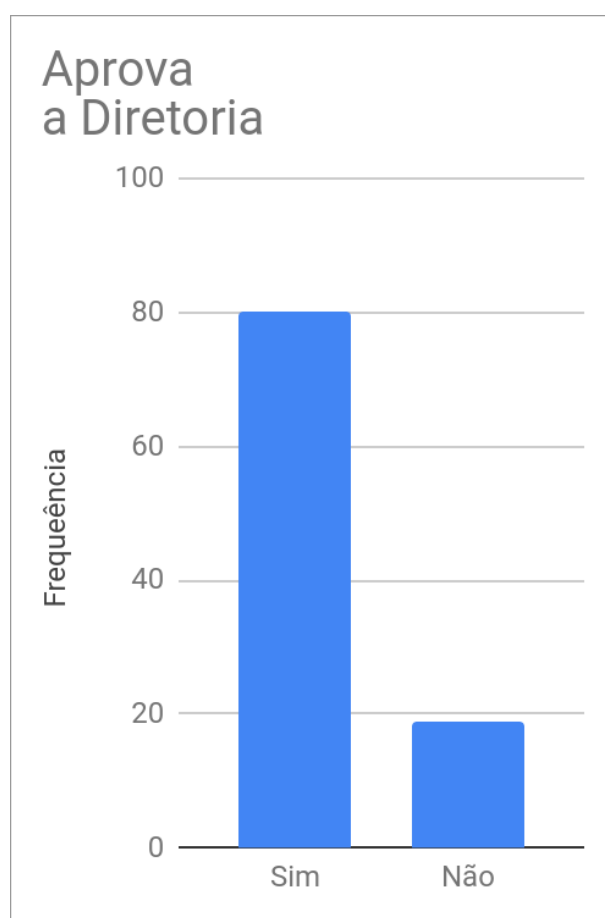
Elaboração própria

Figura 5 – Histograma para a variável “Oportunidade de promoção”



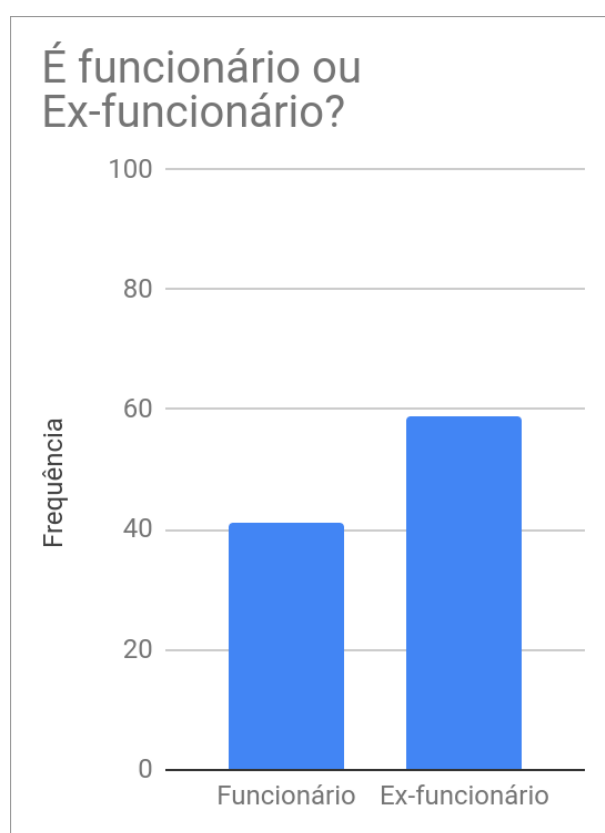
Elaboração própria

Figura 6 – Histograma para a variável “Aprova a diretoria”



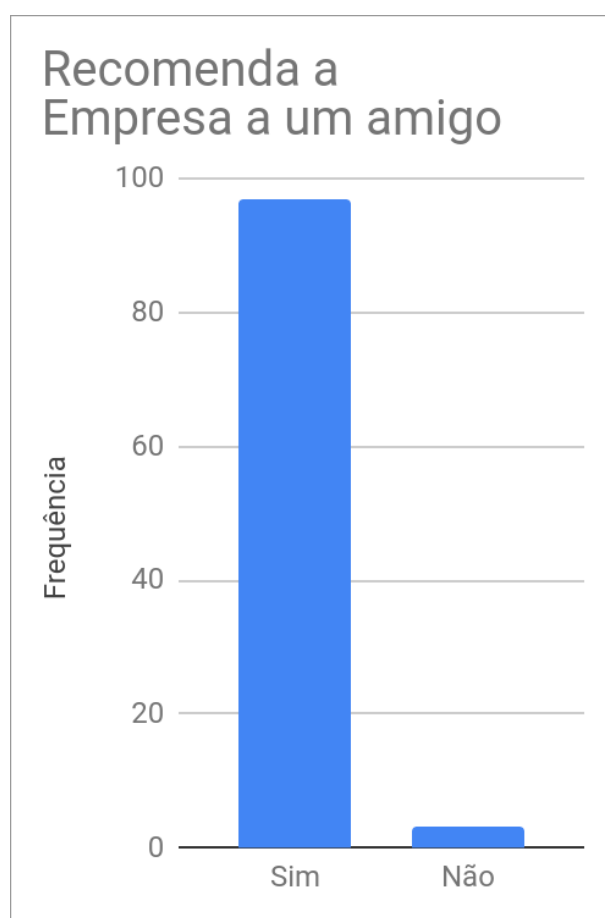
Elaboração própria

Figura 7 – Histograma para a variável “Funcionário”



Elaboração própria

Figura 8 – Histograma para a variável “Recomenda a um amigo”



Elaboração própria

## 4 Conclusão

## Referências

FIELD, A. *Descobrindo a estatística usando o SPSS*. 2. ed. Porto Alegre: Artmed, 2009. Citado na página 8.

JUNIOR, W. T. L. Big data, jornalismo computacional e data journalism: estrutura, pensamento e prática profissional na web de dados. *Estudos em Comunicação*, v. 12, p. 207–222, 2012. Citado na página 8.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. 6. ed. São Paulo: Editora Saraiva, 2017. Citado 2 vezes nas páginas 7 e 8.

PINHEIRO, J. I. D. et al. *Estatística Básica-a Arte de Trabalhar com Dados*. Rio de Janeiro: Elsevier Brasil, 2009. Citado na página 9.

# Glossário

**EAD** Ensino à Distância. 7

**UFABC** Universidade Federal do ABC. 6, 7

**WWW** World Wide Web. 9