Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Eduardo Moreira Araújo
Júlia Garcia Ribeiro
Nathália Oliveira Moreira
Vitória Santos da Silva

Amostra aleatória simples: estudo sobre a proporção de livros com avarias da Biblioteca Central - UnB

Brasília, DF
14 de novembro de 2023

Conteúdo

1 Introdução

A Biblioteca Central da Unb (BCE) desempenha um papel crucial: promover informações para as atividades de ensino e pesquisa da Universidade, por meio de um vasto e rico acervo, atendendo a demanda da comunidade. Entretanto, a baixa qualidade deste vasto acervo pode comprometer a promoção deste acesso fundamental a este patrimonio cultural e acadêmico.

Nesse cenário, surge o interesse em estimar qual a proporção de livros danificados na Biblioteca Central da UnB e quais tipos de danos são mais recorrentes entre eles. Para tal, selecionamos uma amostra aleatória simples sem reposição de uma secção única da BCE para investigar esta situação. A secção de interesse é indicada pela Classe 0, que trata sobre generalidades, informação e organização. Este projeto é um desdobramento do trabalho de amostragem "Estudo sobre a qualidade física dos livros da BCE" realizado no primeiro semestre de 2023, na disciplina Técnicas de Amostragem, pelos alunos Lucas Coelho Christo Fernandes, Luiz Gustavo Jordão Graciano e Raissa Alvim Teixeira;

Espera-se com esse estudo, não só obter um relatório sobre o estado dos livros da biblioteca, mas também, aumentar o conhecimento da administração da BCE em relação aos seus livros para que seja reavaliada o procedimento em relação a preservação desse patrimônio, melhore a satisfação dos frequentadores da biblioteca e contribua de maneira geral com o ambiente.

A parte principal do trabalho está dividida em Metodologia, Resultados e Considerações Finais. Na metodologia explicamos como o desenho do plano amostral para seleção de livros foi esquematizado e definimos as técnicas estatísticas utilizadas, tais como: teste do Qui-Quadrado para comparação de proporções, regressão logística e análise de discriminante. Além disso, discutimos o processo de estimação por amostra aleatória simples, no caso em que temos proporções. Nos resultados realizamos uma análise descritiva dos dados e relatamos o que foi encontrado a partir das técnicas citadas no contexto metodológico. Por fim, nas considerações finais fazemos o resumo dos resultados e indicamos limitações da análise e sugestões para estudos futuros. Em anexo apresentamos o código do *software SAS* utilizado para a execução das análises estatísticas.

2 Metodologia

2.1 Plano amostral

O método de amostragem usado para a elaboração deste trabalho foi a amostra aleatória simples sem reposição. Nela todas as amostras distintas possíveis possuem a mesma probabilidade de seleção. Além disso, cada um dos elementos não sorteados possuem a mesma chance de serem selecionados para a amostra (Bolfarine e Bussab, 2005). Para tal, os elementos da população foram enumerados e sorteados aleatoriamente usando tabelas do livro "A Million Random Digits with 100000 Normal Deviates" do autor George Johnson.

As estantes da classe 0 foram enumeradas de 1 a 5 e designadas para cada um dos estudantes do grupo usando a página destinada a cada um, começando pela primeira linha e primeira coluna e pulando os números que foram sorteados anteriormente por alguém do grupo. Da página 5, foi selecionada a estante 1 (Eduardo); da pagina 12, a estante 3 (Vitória); da pagina 6, a estante 5 (Júlia) e por fim da pagina 1, a estante 2 (Nathália), após o numero 1 e 3 serem pulados por já pertencerem a outro integrante do grupo.

Figura 1: Fotografia das estantes da secção 0

O N de cada estante foi contado manualmente e definido como valor máximo que a enumeração do livro poderia valer. Foram coletadas 25 amostras de cada estante, usando novamente as páginas do livro de George Johnson (primeira linha e primeira coluna), selecionando valores de 4 dígitos e descartando os valores repetidos. Em suma, optou-se por contar primeiramente quantos livros existiam em cada estante, para depois selecionar a amostra de forma aleatória. Caso fizéssemos a estimativa de $250\,x\,20 = 5000$, teríamos uma quantidade significativa de valores *missing*. Isso prejudicaria a precisão das análises feitas no trabalho por causa da perda de observações. Cabe ressaltar que não houve recontagem, uma vez que no processo de

contagem mapeamos quantos livros existiam em cada uma das 6 divisões das 20 prateleiras existentes. O processo durou, em média, 2 horas para cada integrante.

Amostra selecionada com $N_1 = 4393$ (Eduardo):
1961, 2784, 3011, 3656, 2665, 3001, 2759, 1417, 2800, 0642, 3873, 2587, 1794, 3134, 0262, 2836, 0346, 2744, 3423, 1878, 1229, 0645, 3253, 1107, 3771.

Amostra selecionada com $N_2 = 3151$ (Nathália):
1009, 2749, 2063, 0200, 1665, 0842, 2689, 2320, 0336, 0699, 0190, 2529, 0937, 1538, 3113, 1165, 1280, 1573, 2366, 1217, 1768, 0657, 2768, 1992, 0108.

Amostra selecionada com $N_3 = 2663$ (Vitória):
0347, 2597, 0433, 0545, 1099, 0909, 0181, 0457, 0782, 1193, 0676, 2545, 1206, 0115, 2226, 1887, 0636, 0730, 0944, 1129, 1713, 0361, 1860, 2440, 1068.

Amostra selecionada com $N_5 = 3584$ (Júlia):
0987, 0336, 2596, 2786, 0462, 1407, 2523, 1701, 3416, 0264, 2763, 0473, 0206, 0786, 0400, 2596, 0165, 1845, 1610, 0222, 2147, 3350, 3514, 0577, 2310.

Para cada livro foi identificado seu endereço na biblioteca, o seu numero de exemplar, o ano de impressão, sua idade, se existe algum defeito ou avaria e, caso tenha, de qual tipo. As avarias para esse trabalho foram categorizadas em 3 tipos: 1 - Dano na capa, 2 - Oxidação ou costura do livro aparente e 3 - Marcação por lápis ou caneta.

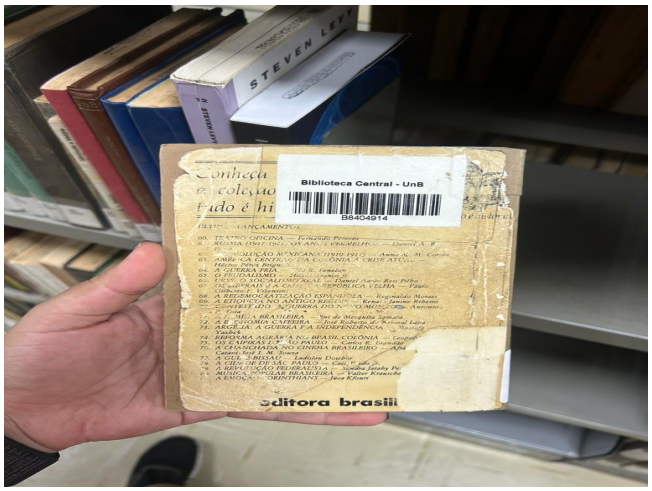


Figura 2: Avaria tipo 1, dano na capa.

Figura 3: Avaria tipo 2, Oxidação ou costura do livro aparente.

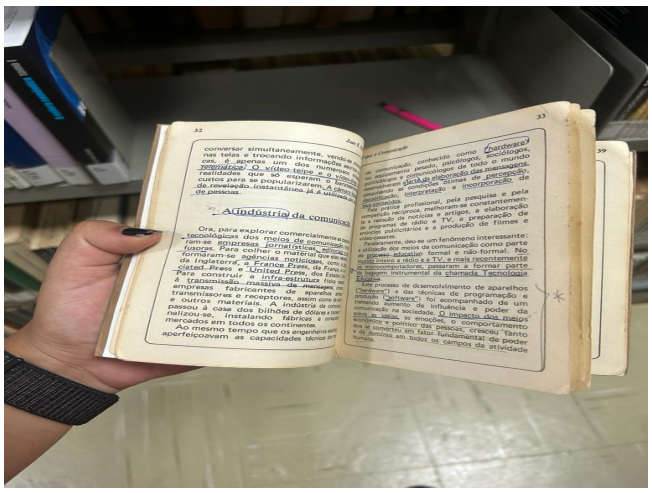


Figura 4: Avaria tipo 3, marcação por lápis ou caneta.

2.2 Estimação

Com base nas amostras coletadas, foi estimado os valores de livros com avarias por meio do estimador de proporção. Para isso, foi usada a fórmula:

$$\hat{p} = \frac{y}{n}$$

Onde $\hat{p}$ é o estimador da proporção, y é o número de elementos na amostra selecionada que possuem avarias e n é o tamanho da amostra coletada, nesse caso, 100. Para o calculo do y refente a n, os livros foram classificados em uma tabela usando a legenda 0, para livros sem a existência de avaria e 1, para livro com a existência de pelo menos um dos tipos de avarias citadas.

O estimador de variância também foi calculado usando os mesmos valores utilizados no estimador de proporção usando a seguinte fórmula:

$$\widehat{Var}(p) = (1 - f)\frac{pq}{n - 1}$$

Como o N total da população não foi possível de ser obtido, o estimador da variância calculado sofreu uma leve modificação. Nesse sentido, sabendo que $f = \frac{n}{N}$, o termo (1 - f) será excluído do cálculo, visto que a amostra representa apenas uma pequena fração do total de livros e não temos informações precisas sobre o tamanho exato da população dos exemplares da Classe 0 na Biblioteca e por tanto a correção aplicada teria uma influência baixa no resultado final (Cochran,1977). É importante ressaltar que isto ocorreu pois a estante 4 não foi sorteada por nenhum dos estudantes do grupo durante as consultas nas tabelas aleatórias. A variância, então, foi dada pela seguinte fórmula:

$$\widehat{Var}(p) = \frac{pq}{n - 1}$$

2.3 Teste do Qui-Quadrado

Uma hipótese levantada no estudo é de que, independente da fileira onde os livros estão localizados, a proporção de obras com avarias não se altera significativamente, ou seja, as proporções de livros defeituosos podem ser consideradas iguais entre as quatro amostras. Nesse sentido, precisamos de um teste estatístico para validar ou rejeitar tal afirmação.

Considerando que estamos trabalhando com uma variável resposta qualitativa nominal que nos indica se o livro possui ou não avaria e a variável explicativa também é categorizada e representa qual a estante do livro, então, precisamos medir a associação entre duas variáveis qualitativas.

O objetivo do Teste do Qui-quadrado para comparar proporções é verificar a igualdade de proporções de uma determinada variável categórica nas diversas amostras, representadas pelas categorias da outra variável. Como os tamanhos das amostras das categorias da variável "Fileira/Estante" foram fixadas a priori ($n = 25$), tem-se o denominado Teste de Homogeneidade (Bussab e Morettin, 2003).

Considerando que p_i é a proporção de avarias em cada uma das quatro fileiras, temos as seguintes hipóteses:

$$\begin{cases} H_0 : p_1 = p_2 = p_3 = p_4 \\ H_1 : p_i \neq p_j, \text{ para algum } i \neq j \end{cases}$$

A estatística utilizada nesse teste é dada pela fórmula:

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- e_{ij} = valor esperado na i-ésima linha e na j-ésima coluna e é dado por:

$$\frac{(total\ da\ linha\ i) \times (total\ da\ coluna\ j)}{total\ geral}$$

- o_{ij} = valor observado na i-ésima linha e na j-ésima coluna

Portanto, considerando a hipótese nula (H_0) verdadeira, a estatística χ^2 seguirá uma distribuição Qui-Quadrado com $v = (4-1)(2-1) = 3$ graus de liberdade, em que 4 representa o número total de linhas da tabela de contingência e 2 o número total de colunas. Cabe ressaltar que os pressupostos para utilização do teste foram atendidos: independência entre amostras, no máximo de 20% das frequências esperadas inferiores a 5 e tamanho amostral adequado (Agresti, 2007).

2.4 Regressão logística: definições e fórmulas

A análise de regressão logística binária é um instrumento eficaz para verificar a relação entre duas ou mais variáveis no caso específico em que a resposta (Y) é dicotomizada em "sucesso" ($Y = 1$) e "fracasso" ($Y = 0$). Para o estudo em questão, sucesso envolve o livro possuir algum tipo de avaria e fracasso representa o livro em perfeitas condições. Sua modelagem é feita a partir da equação:

$$P(Y_i = 1|X_{1i}, \ldots, X_{pi}) = \pi(X_i) = \frac{e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}}$$

em que a probabilidade de sucesso da variável resposta ($Y = 1$) está em função das variáveis explicativas X_i, $i = 1, 2, \ldots, p$.

Tal equação pode ser escrita de maneira linear pela transformação *logito*:

$$\pi^*(X_i) = ln\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$$

O parâmetro β_j corresponde ao efeito do aumento de uma unidade de X_j sobre o logaritmo neperiano da chance de sucesso ($Y = 1$), mantendo as demais variáveis constantes. Dessa forma, e^{β_j} tem como efeito a multiplicação na *odds* (chance de sucesso) para o aumento de uma unidade de X_j, mantendo as demais variáveis constantes (Agresti, 2007).

Para testar se há ausência de regressão, consideramos os testes da razão de verossimilhança, Wald e Score para avaliar conjuntamente se os coeficientes de regressão são iguais ou diferentes de zero (teste de significância), ou seja:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i = 1, ..., p. \end{cases}$$

Por outro lado, também testamos individualmente cada um dos coeficientes de regressão para verificar os efeitos das variáveis explicativas (teste de significância), logo, temos a seguinte configuração:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0, \text{ para } i = 1, ..., p \end{cases}$$

No contexto deste trabalho, as variáveis explicativas que serão avaliadas são “Idade do livro” e “Fileira/Estante do livro”. Nessa perspectiva, vamos verificar se a probabilidade de uma obra apresentar algum defeito é influenciada por alguns desses fatores. Se exister associação, conseguiremos quantificar o poder de influência.

2.5 Classificação por regressão e discriminante linear

Uma etapa interessante do trabalho consistiu em tentar classificar os livros em "com avaria" e "sem avaria". O primeiro método de classificação levou em consideração as probabilidades estimadas no modelo de regressão logístico em que a idade do livro é um fator significativo. Nesse sentido, uma probabilidade específica foi definida como critério de alocação das categorias: $\pi_1 > 0,5 = 1$ e $\pi_1 \leq 0,5 = 0$, por exemplo. O resultado final é uma matriz de confusão, caracterizada como uma ferramenta que nos indica a classificação dos livros nas classes reais e as preditas pelo modelo. Nessa perspectiva, conseguimos verificar quantas classificações certas e erradas obtivemos em todas as combinações possíveis.

Em um segundo momento, utilizamos uma técnica multivariada denominada análise de discriminante linear. De acordo com Johnson e Wichern (2007), trata-se de uma ferramenta para alocar objetos em duas ou mais classes (populações) de acordo com regras estabelecidas e avaliar a qualidade da alocação. Em suma, devemos encontrar regras que indicam as diferenças entre objetos provenientes de diferentes populações conhecidas (técnica supervisionada).

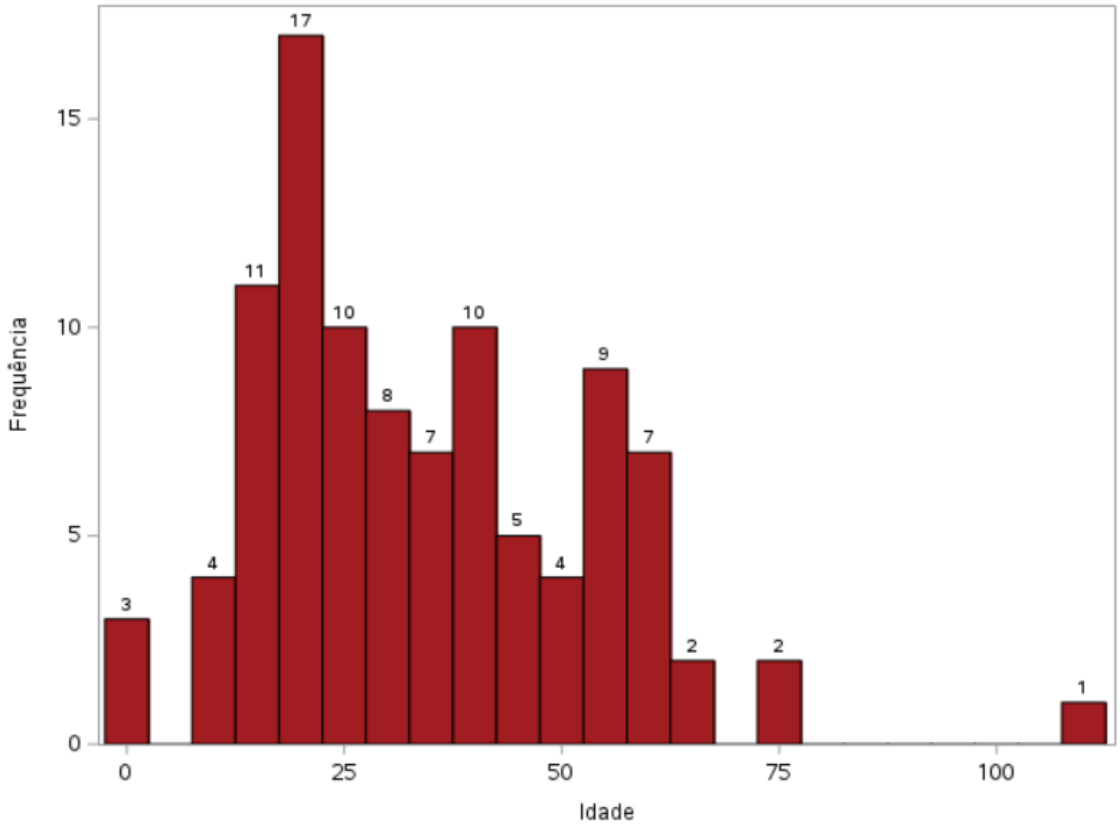
Com a ajuda da *PROC DISCRIM* do software SAS, tentamos encontrar funções lineares que maximizam a separação entre as médias de idades dos dados observados para os livros com e sem avarias. Novamente, o resultado foi uma matriz de confusão para avaliar a qualidade da classificação. Por fim, também tentamos expandir a análise para as quatro possíveis categorias de defeitos: 0 - sem defeitos, 1 - dano na capa, 2 - oxidação ou costura do livro aparente e 3 - marcação por lápis ou caneta.

Obs: Para a execução da regressão logística e da classificação por discriminante linear, os dados foram separados aleatoriamente em dois grupos: treinamento e validação. Nesse sentido, a concepção é verificar se os resultados e interpretações permanecem robustos e se encontram na mesma direção em ambos os casos. Um problema evitado, nesse sentido, é o de *overfitting*, por exemplo.

3 Resultados

3.1 Análise Descritiva

Figura 1: Idade dos Livros



3.1.1 Idade dos exemplares

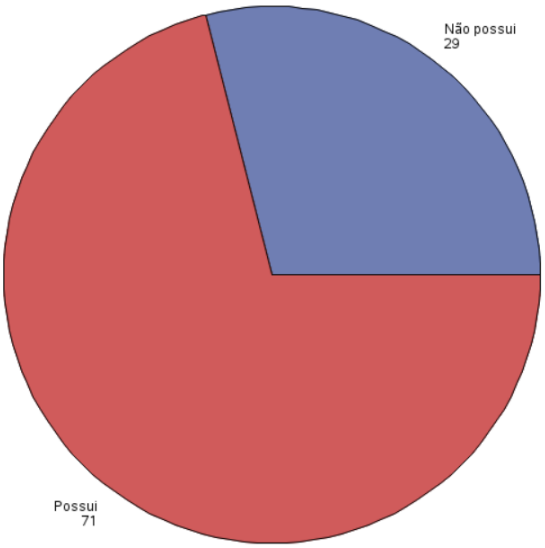
Quadro 1: Medidas Descritivas para a variável Idade dos Livros

Medidas	Resultado
Média	34,11
Mínimo	1
Máximo	110
1º Quartil	20
3º Quartil	47,50
Desvio Padrão	18,57

De acordo com o histograma das Idades dos livros, é notável que a maioria dos exemplares têm idade entre 24 e 40 anos. Apresentando média em torno de 34 anos. É possível observar também que há outliers, tendo 3 livros com cerca de 1 ano e também uma obra com excepcionais 110 anos. É perceptível que 50% dos dados centrais estão em torno de 20 e 47,50 anos.

3.1.2 Presença de avaria

Figura 2: Se possui Avaria

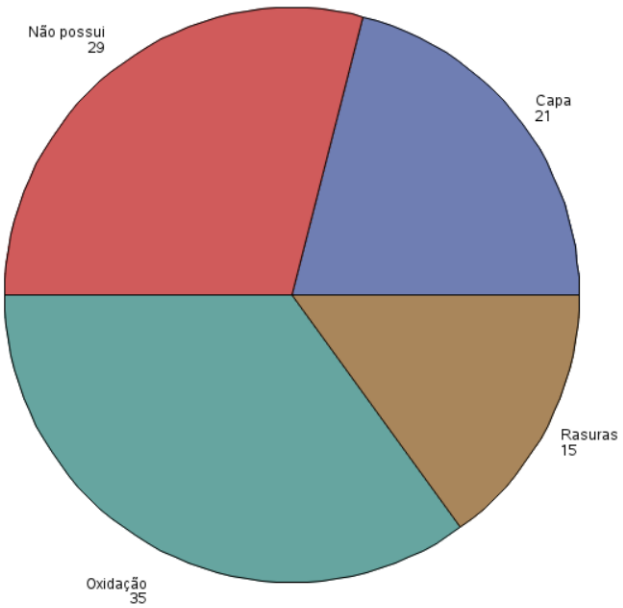


Observando-se a Figura 2 nota-se que a maioria dos livros da amostra apresentam avarias e 29% dos livros não apresentaram nenhum defeito. Portanto, a proporção de avarias estimada do estudo é de 0,71 (71%).

3.1.3 Tipos de avarias

Como foi dito na introdução deste estudo, as avarias observadas nos exemplares são três: se o livro apresenta uma capa danificada, se há oxidação em suas folhas e se há rasuras ou anotações em seu interior. Dada esta informação, estes foram os resultados observados:

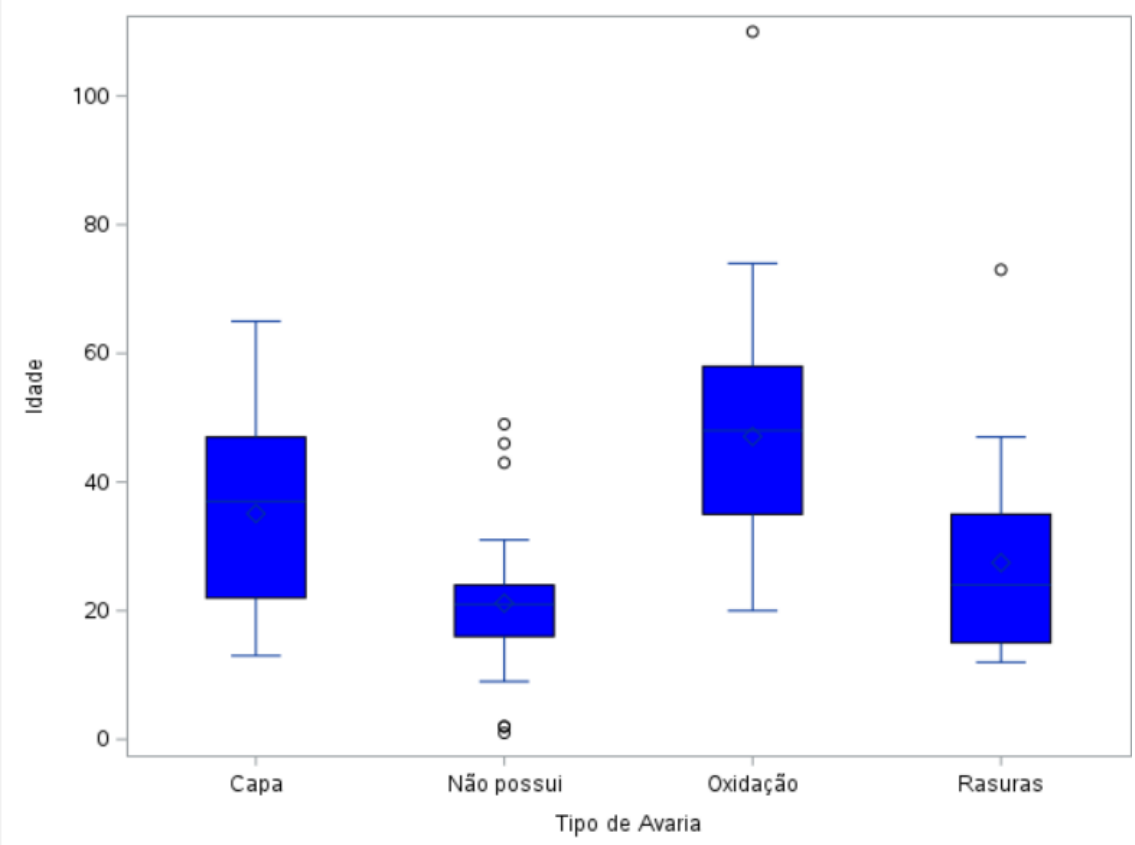
Figura 3: Tipo de Avaria



Nota-se que dentre os tipos de avarias observados, a maior quantidade de casos foi para oxidação das páginas, seguido por 21 exemplares que apresentaram dano na capa e o caso em que houveram menos observações desta avaria são as rasuras, contando com 15 casos observados.

3.1.4 Tipo de avaria por idade do livro

Figura 4: Tipo de Avaria por Idade do Livro



Na Figura 4, observa-se os tipos de avarias e a ocorrência destas de acordo com a idade do livro. Nota-se que a capa danificada e a oxidação das páginas são avarias típicas encontradas em livros com uma idade mais avançada, podendo ser resultado da ação do tempo. Enquanto casos de rasuras ou de não possuir avarias são características mais comuns entre livros com menor idade.

3.2 Variância e intervalo de confiança

A partir desta seção vão ser trabalhadas questões inferenciais do estudo e portanto, iniciaremos com a estimação da variância de $\hat{p}$. A princípio, foi dito que a amostragem aleatória simples sem reposição foi feita no estudo. Logo, as estimações serão feitas com base nesta informação, considerando o caso de estimação para proporção.

Antes de iniciar o cálculo é importante lembrar que será calculada a variância estimada de $\hat{p}$ e não a variância de p, isso se deve pelo fato de não se ter o os valores totais da população, como a proporção total de livros com e sem avarias na Biblioteca.

$$\widehat{Var}(p) = (1 - f)\frac{pq}{n-1}$$

com

$$f = \frac{n}{N}$$

Como a amostra é pequena em relação ao total de livros e o tamanho populacional exato não é conhecido para os exemplares da Classe 0 da Biblioteca, o fator de correção terá um efeito muito pequeno e portanto, de baixo impacto no resultado final. Por este motivo, (1 - f) será retirado do cálculo. Logo, o cálculo será da seguinte forma:

$$\widehat{Var}(p) = \frac{pq}{n-1}$$

Sabendo-se que a proporção de livros com avarias (0,71) e sem avarias (0,29) são conhecidas a partir do processo de estimação, e que há 100 livros na amostra deste estudo, tem-se o seguinte resultado:

$$\widehat{Var}(p) = \frac{0,71.0,29}{100-1} = 0,002079$$

Agora será calculado o intervalo de confiança para a este estimador, para que se possa entender a variabilidade do estimador da proporção de avarias. Esse intervalo segue a seguinte fórmula:

$$IC = \widehat{p} \mp Z_{\alpha/2}\sqrt{\widehat{Var}(p)}$$

Para a fórmula acima deve-se considerar o $\hat{p}$ como o estimador da proporção de livros que apresentam as características de interesse do estudo e o Z é o quantil da distribuição normal considerada para o estudo, que será o padrão de 1,96 ($\alpha = 0.05$). A normal foi considerada porque tem-se uma amostra com n > 30. Por fim, haverá a multiplicação de Z pelo erro padrão do estimador, que é dado pela raiz da variância estimada. Substituindo os valores:

$$IC = 0,71 \mp Z_{0,025}\sqrt{0,002079} = (0,6206; 0,7993)$$

Olhando o resultado acima, é possível dizer que, com uma confiança de 95%, a real proporção de livros com avarias da Classe 0 da Biblioteca está contida no intervalo acima, ou seja, está entre 0,62 e 0,80. O estimador $\hat{p}$ da amostra deste estudo está dentro do intervalo, sendo igual a 0,71 (estimativa pontual).

3.3 Comparação entre proporções

Uma preocupação importante consiste em saber se a proporção de livros defeituosos é homogênea em cada uma das quatro fileiras selecionadas aleatoriamente. Nesse sentido, precisamos verificar pelo teste do Qui-Quadrado de homogeneidade se as proporções podem ser consideradas iguais ou não. Vejamos a tabela abaixo:

Tabela 1: Distribuição do número de livros segundo a existência de avaria e fileira selecionada - Brasília - 2023

Fileira	**Avaria**		**Total**
	Sim	Não	
1	14 (56%)	11 (44%)	25
2	18 (72%)	7 (28%)	25
3	21 (84%)	4 (16%)	25
5	18 (72%)	7 (28%)	25
Total	71	29	100

A hipótese inicial, pensada antes do estudo, era que a proporção de livros com avarias independe das estantes ou fileiras onde os livros estão localizados, ou seja, os livros da classe 0 tenderiam a ser homogêneos. No momento da coleta dos dados, percebeu-se que essa afirmação precisava ser melhor investigada. Conforme verificamos na tabela acima, a proporção de livros defeituosos na fileira 1 aparenta ser diferente das demais, sobretudo, da fileira 3 que apresenta a estimativa de 84%. A primeira fileira é composta por livros de tamanho médio e a terceira por obras bem maiores, o que poderia caracterizar um fator de influência para a presença ou não de falhas, por exemplo.

O valor obtido para a estatística do teste Qui-Quadrado foi $\chi^2 = 4,8082$, considerando 3 graus de liberdade. O p-valor obtido foi de aproximadamente $0,186$. Portanto, com um nível de significância dado por $\alpha = 0,05$, não rejeitamos a hipótese nula de homogeneidade na variável resposta "Avaria" considerando as categorias da variável explicativa "Fileira". Em suma, não existem evidências para afirmarmos que o número ou a proporção de avarias difere significativamente nas quatro fileiras amostradas. Embora tenhamos verificado variação nos tamanhos dos livros, isso não foi o suficiente para desenvolver uma heterogeneidade na comparação de livros com avarias por fileira.

3.4 Regressão logística: aplicação

Um dos objetivos do trabalho consistiu em identificar, além da existência de avarias nos livros, algumas variáveis que poderiam exercer influência sobre o fato do livro apresentar ou não uma falha.

A fileira que cada um dos integrantes do grupo selecionou aleatoriamente foi uma das variáveis investigadas. No tópico anterior não rejeitamos a hipótese de igualdade entre as proporções de avarias considerando as fileiras, ou seja, tal variável não parece ser significativa para explicar a quantidade de livros defeituosos. Na regressão logística, podemos novamente verificar se tal cenário é mantido.

A segunda variável de interesse é a idade do livro. A expectativa é que quanto mais velho um livro, maior é a probabilidade de que a obra apresente avaria. Nesse sentido, se faz necessário quantificar tais probabilidades e a razão de chances (odds ratio). Vejamos na tabela abaixo os resultados para os testes de significância conjunta dos coeficientes de regressão:

Tabela 2: Testes globais de significância para ausência de regressão logística

Teste	**Qui-Quadrado**	**Graus de liberdade**	**P-valor**
Razão de Verossimilhança	29,72	4	<.0001
Score	23,90	4	<.0001
Wald	18,12	4	0,0012

Considerando os valores obtidos, verificamos que a hipótese nula de ausência de regressão será rejeitada com um nível de significância de 5%. Portanto, é possível ajustar um modelo logístico com os dados. Para verificar quais variáveis cabem ou não no modelo, vamos verificar os testes individuais para os coeficientes de regressão, conforme tabela abaixo:

Tabela 3: Testes individuais de significância para os coeficientes de regressão

Parâmetro	**GL**	**Estimativa**	**Erro Padrão**	**Qui-Quadrado**	**P-valor**
Intercepto	1	-2,1751	0,7493	8,4270	0,0037
Idade	1	0,0860	0,0217	15,6940	<.0001
Fileira 2	1	0,9047	0,6821	1,7595	0,1847
Fileira 3	1	1,2492	0,7863	2,5239	0,1121
Fileira 5	1	0,2726	0,7213	0,1428	0,7055

Observamos que o intercepto é diferente de 0 e o coeficiente de regressão para a variável “Idade” é significativo. Além disso, não rejeitamos a hipótese nula em que os coeficientes das fileiras são iguais a zero.

Novamente podemos depreender que as estantes não exercem influência significativa sobre a variável reposta "Avaria". Esse resultado é condizente com o que observamos na etapa de comparação entre as proporções, embora o processo tenha sido diferente.

Por outro lado, um modelo com a idade do livro parece fazer bastante sentido. Retirando as fileiras e reajustando o modelo final, temos a seguinte configuração para a equação com o intercepto e a variável "Idade":

$$logito = -1,61 + 0,087 * idade$$

Com 95% de confiança, a razão de chances associada a idade do livro está contida no intervalo entre $1,045$ e $1,138$. A estimativa pontual foi de $1,091$ ($e^{0,0868}$). Nesse sentido, para cada ano que aumenta na idade do livro a chance estimada de sucesso (odds) é multiplicada por valores entre $1,045$ e $1,138$. Em suma, existe um aumento na chance do livro apresentar defeito conforme cada ano é acrescido, o que também influencia na probabilidade de falha.

Tabela 4: Probabilidades estimadas de avaria segundo a idade dos livros em anos completos - Brasília - 2023

Idade do Livro (anos)	**Probabilidade Estimada**
1	0,179
10	0,323
20	0,532
30	0,730
40	0,865
50	0,938
60	0,973

Por fim, realizamos o teste de Adequabilidade de Ajustamento de Hosmer e Lameshow. A hipótese nula consiste em afirmar que o modelo de regressão logística ajusta-se aos dados. Trata-se de observar a qualidade de ajuste do modelo a partir dos valores preditos das probabilidades. Com um nível de significância de 5%, não rejeitamos a hipótese nula e consideramos que o modelo está bem ajustado.

Tabela 5: Teste de Adequabilidade de Ajustamento (Hosmer e Lameshow)

Qui-Quadrado	**GL**	**P-valor**
13,87	8	0,085

3.5 Classificação dos livros

A partir da regressão logística e o fato da variável "Idade" ter sido considerada significativa, surgiu o seguinte questionamento: conseguimos desenvolver um processo capaz de classificar se um livro tem ou não avaria com o auxílio da variável idade?

Se usarmos as probabilidades estimadas da regressão logística e considerarmos que a classificação de "sem avaria" será dado para observações com probabilidade abaixo ou igual a $0,5$ e a classificação de "com avaria" para observações com probabilidades maiores que $0,5$, teríamos a seguinte matriz de confusão:

Tabela 6: Matriz de confusão - regressão logística com $\pi_1 > 0,5$

Real \Predito	**0**	**1**
0	10	19
1	08	63

Observamos que a taxa de acertos foi de 73%, entretanto, houve uma quantidade significativa de erros no que tange livros sem avarias que foram classificados como defeituosos. Se formos mais rigorosos, ou seja, a classificação de "com avaria" passa a ser para observações com probabilidades maiores que $0,6$, então teríamos um aproveitamento de 75% com menos erros do caso anterior, porém, com mais erros em classificar os livros com avarias. Vejamos esse cenário na tabela abaixo:

Tabela 7: Matriz de confusão - regressão logística com $\pi_1 > 0,6$

Real \Predito	**0**	**1**
0	20	09
1	16	55

Uma outra forma de pensarmos o problema é utilizando uma função que faz o papel de discriminante linear. Com as prioris de $\pi_0 = 0,4$ e $\pi_1 = 0,6$ conseguimos uma matriz de confusão muito similiar ao obtido no caso anterior, conforme observamos a seguir:

Tabela 8: Matriz de confusão - discriminante linear com prioris predefinidas

Real \Predito	**0**	**1**
0	22	07
1	18	53

Por fim, podemos ampliar o nível de classificação que utilizamos até aqui. Se a categoria da variável resposta que indica avaria do livro fosse ampliada para os três

tipos de defeitos (1: estado da capa; 2: oxidação das páginas e/ou costura do livro aparente; 3: uso de marca textos ou canetas ou lápis), a seguinte matriz de confusão será gerada:

Tabela 9: Matriz de confusão - discriminante linear com prioris proporcionais

Real \Predito	**0**	**1**	**2**	**3**
0	26	00	03	00
1	09	00	12	00
2	07	00	28	00
3	11	00	04	00

Nessa perspectiva, a classificação dos itens sem avarias teve uma taxa de acertos de quase 90%. Apesar disso, 27% dos livros possuem algum tipo de defeito, mas foram considerados sem avarias. Por outro lado, houve uma classificação satisfatória de 80% para os livros que apresentam oxidação das páginas e/ou costura do livro aparente, o que indica um efeito maior da variável idade sobre esse tipo específico de evento.

O fato negativo reside na classificação dos livros com avarias dos tipos 1 e 3. A taxa de erro foi de 100% considerando o total de 36 observações. Nesse caso, uma outra função discriminante poderia ser ajustada ou a melhoria na classificação passa pela inclusão de um número maior de variáveis explicativas no modelo. A taxa de acertos geral foi de 54%.

Em síntese, as classificações por regressão logística ou com o uso de uma função discriminante linear nos retornam resultados relativamente satisfatórios, considerando que utilizamos apenas uma variável explicativa na modelagem dos dados.

4 Considerações Finais

A proporção estimada de livros com avarias para a secção sobre generalidades, informação e organização foi de 0,71. Considerando o intervalo de confiança de 95%, o real valor do parâmetro reside entre 0,62 e 0,80. Nesse sentido, há um número elevado de livros com avarias, sejam tais defeitos por causa do estado da capa, oxidação ou costura aparente, ou por rasura por lápis e caneta.

A hipótese de igualdade entre as proporções de avarias considerando cada uma das quatro estantes selecionadas não foi rejeitada pelo teste do Qui-Quadrado. Nesse sentido, consideramos que não existe associação entre a localização do livro (estante/fileira) e a presença de avaria nos exemplares. Na modelagem por regressão logística reafirmamos tal cenário. Nesse sentido, a probabilidade de um livro apresentar falha não é influenciada significativamente pela localização da obra.

Considerando a regressão logística, verificamos que a idade é um fator significativo. Em síntese, para cada ano que aumenta na idade do livro a chance estimada de sucesso (apresentar falha) é multiplicada por valores entre $1,045$ e $1,138$. Portanto, existe um aumento na chance do livro apresentar defeito conforme cada ano é acrescido, o que também influencia na probabilidade de falha.

Foi possível classificar os livros a partir da própria regressão logística ou com o uso de uma função linear discriminante. Os melhores resultados apresentaram uma taxa de acertos de 75%, sendo este um valor satisfatório se consideramos que apenas a variável explicativa idade foi utilizada. Quando expandimos a análise, verificamos que foi possível classificar bem os livros sem avaria e com avarias do tipo oxidação, o que corrobora para a existência de uma maior associação entre a idade e o fenômeno de livros oxidados, em comparação com os outros tipos de defeitos.

A amostragem por amostra aleatória simples sem reposição se mostrou interessante na medida em que nos retornou uma quantidade significativa de informação e muitas técnicas estatísticas foram utilizadas. Apesar disso, uma limitação do estudo é que não foi possível identificar o tamanho total da população, sobretudo, porque uma das fileiras não foi sorteada aleatoriamente. Caso essa estante seja heterogênea em relação as demais, algumas interpretações, na realidade, podem ser diferentes do que obtivemos neste projeto. Além disso, uma amostragem por *cluster* poderia ser pensada como uma forma mais adequada para obtenção dos dados.

Por fim, uma sugestão que pode ser indicada é a obtenção de mais variáveis explicativas para a modelagem dos dados, além da idade do livro. Nesse sentido, se existir um conjunto maior de covariáveis que ajudam a identificar se um livro possui ou não avaria, bem como o tipo de defeito, então, com uma amostra é possível encontrar um algoritmo de classificação dos livros. Se o processo for eficiente e com uma taxa de erros baixa, então, toda a base de dados referente aos livros da BCE poderia ser classificada posteriromente com alta precisão.

5 Referências Bibliográficas

AGRESTI, Alan. An Introduction to Categorical Data Analysis, 2a ed.. New York: John Wiley & Sons, 2007

BOLFARINE, H. e BUSSAB, W. O. Elementos de Amostragem. Edgard Blucher. 2005.

BUSSAB, W. e MORETTIN, P. , Estatística Básica, 7a edição. São Paulo: Ed. Saraiva, 2003.

COCHRAN, W. G. Sampling Techniques. 3º edition. Wiley. 1977.

R. A. JOHNSON and D. W. WICHERN. Applied multivariate statistical analysis. Prentice Hall, 2007. Sexta Edição.

6 Anexo

Para acessar o código bruto em formato SAS da análise descritiva e dos resultados, acessar link abaixo:

Código em SAS.

Caso não seja de interesse baixar o código na íntegra, basta visualizá-lo logo abaixo.

```
/* ########################################################################## */
/* ######## ANÁLISE DESCRITIVA - TRABALHO AMOSTRAGEM ######## */
/* ########################################################################## */


/* Importar banco de dados */

proc import datafile="/home/u59041777/trab_avaria (1).xlsx"
        out=banco
        dbms=xlsx replace;
run;


/* Histograma */

proc sgplot data=banco;
  histogram Idade /
    binwidth=5       /* Largura dos bins */
    fillattrs=(color="#A11D21") /* Cor das barras do histograma */
    datalabel;       /* Exibir rótulos de dados */

  xaxis label='Idade'; /* Rótulo do eixo X */
  yaxis label='Frequência'; /* Rótulo do eixo Y */
run;

/* Gráfico de Setores Avarias */

data banco1;
  set banco;
  if Avaria = 0 then Avaria_Label = 'Não possui';
  else if Avaria = 1 then Avaria_Label = 'Possui';
run;


proc gchart data=banco1;
  pie Avaria_Label;
run;

/* Gráfico de Setores Tipos de Avarias */

data banco2;
  set banco;
  if Tipo_avaria = 0 then Tp_Avaria_Label = 'Não possui';
  if Tipo_avaria = 1 then Tp_Avaria_Label = 'Capa';
  if Tipo_avaria = 2 then Tp_Avaria_Label = 'Oxidação das folhas';
  else if Tipo_avaria = 3 then Tp_Avaria_Label = 'Rasuras';
run;
```

```
proc gchart data=banco2;
  pie Tp_Avaria_Label;
run;

/* Ano com tipos de avarias */

proc sgplot data=banco2;
   vbox Idade / category=Tp_Avaria_Label lineattrs=(color=black) fillattrs=(color=blue); /*
Defina a cor desejada aqui */
   yaxis label='Idade';
   xaxis label='Tipo de Avaria';
run;


proc print data=banco1;
run;

/* Medidas descritivas Idade */

proc means data=banco mean min max std p25 p75;
   var Idade;
   output out=media_idade mean=media_idade;
run;

/* Calcular a variância chapéu */

data variancia;
p= 0.71;
q = 0.29;
variancia_chapeu = (p*q)/(100-1);
run;

proc print data=variancia;
run;
```

```
/* ########################################################## */
/* ** PROGRAMA-SAS – RESULTADOS **** */
/* ########################################################## */

OPTIONS LS=80 PS=60 NODATE;

/* LEITURA E DIVISÃO DOS DADOS */

proc import
 datafile="/home/u59041762/Dados categorizados/trabalho_avaria_BCE.xlsx"
 out= pessoal.dados DBMS = xlsx REPLACE;
 sheet="Dados";
 GETNAMES = YES;
run;

/* Variável de identificação 1 a 100 */
DATA ID1 (drop = i);
    id = 0;
    do i = 1 to 100;
          id + 1;
       output;
          end;
RUN;

DATA pessoal.dados1;
  set pessoal.dados;
  set ID1;
RUN;

/* dados de treinamento */

proc surveyselect data = pessoal.dados1 method = SRS rep = 1
  sampsize = 50 seed = 14112023 out = pessoal.dados_treinamento;
  id _all_;
run;

proc print data = pessoal.dados_treinamento noobs;
run;


/* dados de validação */
data pessoal.dados_validacao;
  set pessoal.dados1;
     if id not in (01,05,07,08,14,15,17,18,21,22,
                   24,25,26,27,28,31,32,36,38,39,
                   41,46,47,49,50,51,53,54,58,59,
                   60,62,64,66,67,69,71,72,76,77,
                   79,81,82,85,86,87,90,91,93,97) then delete;
```

```
run;



/* ---------------------------------------------------------------- */



/* AMOSTRA TREINAMENTO - 50 OBSERVAÇÕES */

/* Regressão Logística - Binária */
proc logistic data = pessoal.dados_treinamento;
 class Fileira /ref=first param=ref;
 model Avaria (event='1') = Idade_Livro Fileira /covb;
run;

/* Estimação - Probabilidade */
proc logistic data= pessoal.dados_treinamento descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro;
 output out=estim p=pi_est ;
run;

/* Listando os valores estimados de Pi */
Proc print data=estim;
var Avaria Idade_Livro _level_ pi_est;
run;

proc sort data = estim presorted out=ordem;
by pi_est;
run;

Proc print data=ordem;
var Avaria Idade_Livro pi_est;
run;

/* Solicita o teste de Hosmer e Lameshow */
/* de Adequabilidade de Ajustamento */
proc logistic data = pessoal.dados_treinamento descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro /lackfit;
run;
```

```
/* Discriminante Linear */
data new_data;
    set pessoal.dados_treinamento;
    keep Avaria Idade_Livro;
run;

proc discrim data = new_data;
  class Avaria;
  priors '0' = 0.4 '1'=0.6;
run;




/* ---------------------------------------------------------------- */




/* AMOSTRA VALIDAÇÃO - 50 OBSERVAÇÕES */

/* Regressão Logística - Binária */
proc logistic data = pessoal.dados_validacao;
 class Fileira /ref=first param=ref;
 model Avaria (event='1') = Idade_Livro Fileira /covb;
run;

/* Estimação - Probabilidade */
proc logistic data= pessoal.dados_validacao descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro;
 output out=estim p=pi_est ;
run;

/* Listando os valores estimados de Pi */
Proc print data=estim;
var Avaria Idade_Livro _level_ pi_est;
run;

proc sort data = estim presorted out=ordem;
by pi_est;
run;

Proc print data=ordem;
var Avaria Idade_Livro pi_est;
run;
```

```
/* Solicita o teste de Hosmer e Lameshow */
/* de Adequabilidade de Ajustamento */
proc logistic data = pessoal.dados_validacao descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro /lackfit;
run;


/* Discriminante Linear */
data new_data;
    set pessoal.dados_validacao;
    keep Avaria Idade_Livro;
run;

proc discrim data = new_data;
  class Avaria;
  priors '0' = 0.4 '1'=0.6;
run;




/* ------------------------------------------------------------------ */




/* AMOSTRA COMPLETA - 100 OBSERVAÇÕES */


proc import
 datafile="/home/u59041762/Dados categorizados/trabalho_avaria_BCE.xlsx"
 out= pessoal.dados DBMS = xlsx REPLACE;
 sheet="Dados";
 GETNAMES = YES;
run;

/* Vendo conteúdo do arquivo SAS */
proc contents data= pessoal.dados varnum;
run;

/* Análise descritiva e teste do Qui-Quadrado */
proc sort data = pessoal.dados presorted out=ordem;
by Fileira;
run;
```

```
proc means data = ordem n mean median
   std var cv maxdec=2;
   var Avaria;
   class Fileira;
   by Fileira;
run;

proc freq data = pessoal.dados order=data;
 tables Fileira*Avaria /nopercent nocol relrisk chisq;
run;

/* Regressão Logística - Binária */
proc logistic data = pessoal.dados;
 class Fileira /ref=first param=ref;
 model Avaria (event='1') = Idade_Livro Fileira /covb;
run;

/* Estimação - Probabilidade */
proc logistic data= pessoal.dados descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro;
 output out=estim p=pi_est ;
run;

/* Listando os valores estimados de Pi */
Proc print data=estim;
var Avaria Idade_Livro _level_ pi_est;
run;

proc sort data = estim presorted out=ordem;
by pi_est;
run;

Proc print data=ordem;
var Avaria Idade_Livro pi_est;
run;

/* Solicita o teste de Hosmer e Lameshow */
/* de Adequabilidade de Ajustamento */
proc logistic data = pessoal.dados descending;
 class Fileira /ref=first param=ref;
 model Avaria = Idade_Livro /lackfit;
run;


/* Discriminante Linear */
data new_data;
```

```
    set pessoal.dados;
    keep Avaria Idade_Livro;
run;

proc discrim data = new_data;
  class Avaria;
  priors '0' = 0.4 '1'=0.6;
run;

data new_data;
    set pessoal.dados;
    keep Tipo_avaria Idade_Livro;
run;

proc discrim data = new_data;
  class Tipo_avaria;
  priors proportional;
run;
```