LISTA 04 - AVALIAÇÃO DE DESEMPENHO DE SISTEMAS

Allyson Ryan

Agosto, 2025

Inferência Estatística, Análise Exploratória de Dados e Regressão Linear

1 Suponha que durante um período de tempo foram testados 28427 capacitores dos computadores de um grande data center. Foi observado que 615 capacitores apresentaram problemas. Calcule o intervalo de confiança para a proporção de defeitos com 99% de confiança.

(1) Proporção amostral $\hat{p}$

(Usa-se a fração observada de defeitos como estimativa pontual da proporção verdadeira.)

$$\hat{p} = \frac{x}{n} = \frac{615}{28{,}427} \approx 0{,}0216335$$

(2) Verificação das condições para aproximação normal

(Aproximação normal para a proporção requer amostra grande e contagens esperadas de sucessos e fracassos ≥ 10.)

$$n = 28{,}427 \quad \text{(suficientemente grande)}, \qquad n\hat{p} = 615 \geq 10, \qquad n(1-\hat{p}) = 27{,}812 \geq 10.$$

(Logo, a aproximação normal é válida.)

(3) Valor crítico para 99% de confiança

(Para IC bilateral de 99%, usa-se $z_{\alpha/2} = z_{0{,}005}$ da tabela normal.)

$$z_{0{,}005} = 2{,}575829$$

(4) Erro-padrão da proporção

(Quantifica a variabilidade amostral de $\hat{p}$ sob repetidas amostragens do mesmo tamanho.)

$$\mathrm{EP}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}0216335 \times 0{,}9783665}{28{,}427}} = 0{,}00086288$$

(5) Margem de erro

(Extensão máxima esperada da flutuação amostral em 99% dos casos.)

$$\mathrm{ME} = z_{0,005} \cdot \mathrm{EP} = 2{,}575829 \times 0{,}00086288 \approx 0{,}0022226$$

(6) Intervalo de confiança de 99%

(Construído pela regra $\hat{p} \pm \mathrm{ME}$ para proporções sob aproximação normal.)

$$\hat{p} \pm \mathrm{ME} = 0{,}0216335 \pm 0{,}0022226 \quad \Rightarrow \quad \boxed{[0{,}0194109,\ 0{,}0238561]}$$

Interpretação em porcentagem

(Expressão equivalente do intervalo em pontos percentuais, útil para comunicação.)

$$\boxed{1{,}94109\% \ \leq \ p \ \leq \ 2{,}38561\%}$$

(Com 99% de confiança, a proporção verdadeira de capacitores defeituosos situa-se nesse intervalo.)

2 Utilizando os dados apresentados na tabela abaixo, realize uma análise exploratória dos dados, proponha uma distribuição teórica para eles e verifique a aderência desses dados à essa distribuição utilizando os métodos Gráfico e Teste KS.

44.5	43.3	41.4	28.9
35.8	38.4	42.3	35.2
33.8	32.8	36.8	40.4
42.8	38.5	41.5	39.1
44.3	38.8	40.5	30.1
42.7	33.2	36.5	38.4
26.9	40	36.5	35
36.7	40.8	41.4	30.4
36.5	31.4	34.2	36.1
30	35.1	39.4	40.1
39.7	33.7	44.1	36.1
31.7	40.7	38.1	42.4
36.4	42.9	39.5	34.7
23.6	39.6	34.4	35.2
41	37.1	40.3	45.2
41.1	36.7	39.1	43.4
50.1	34	37.3	29.1
37	36.7	37.7	29.5
37.2	36.1	44.3	38.3
47.2	30.4	39.1	43.6
34.2	37.4	39.7	43.5
34.2	43	43	32.9
31.8	37.8	44.8	27.5
32	35.7	32.8	40.8
37.1	37	41.2	48.8

Questão 2 — Análise Exploratória, Ajuste de Distribuição e Teste KS

Dados

(A base contém 100 observações numéricas contínuas, representando medidas de desempenho; para inferir um modelo probabilístico adequado, começa-se pela análise descritiva e pelo formato da distribuição.)

1) Análise Exploratória (EDA)

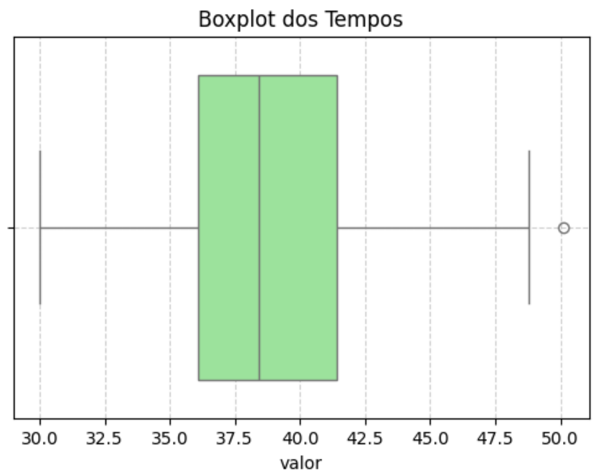
(Resumo descritivo quantifica posição, dispersão e assimetria; isso orienta a escolha de uma família de distribuições candidata.)

$$\begin{aligned}
&n = 100 \quad \textit{(tamanho da amostra)}\\
&\min = 23{,}6, \quad \max = 50{,}1 \quad \textit{(amplitude dos dados)}\\
&\bar{x} = 37{,}70 \quad \textit{(média: tendência central)}\\
&\tilde{x} = 37{,}55 \quad \textit{(mediana: tendência central robusta)}\\
&s = 4{,}897 \quad \textit{(desvio-padrão amostral: dispersão)}\\
&Q_1 = 34{,}625, \quad Q_3 = 41{,}025, \quad IQR = Q_3 - Q_1 = 6{,}40 \quad \textit{(dispersão robusta)}\\
&\text{Assimetria (skew)} \approx -0{,}19 \quad \textit{(leve cauda à esquerda, quase simétrico)}
\end{aligned}$$

(Outliers via regra do IQR: valores fora de $[Q_1 - 1{,}5\,IQR,\ Q_3 + 1{,}5\,IQR]$ indicam pontos atípicos que podem influenciar o ajuste.)

Limite inferior $= 34{,}625 - 1{,}5 \cdot 6{,}40 = 25{,}025$, Limite superior $= 41{,}025 + 1{,}5 \cdot 6{,}40 = 50{,}625$.

(Há um valor abaixo do limite inferior: 23,6; nenhum acima do limite superior. Um único outlier baixo não impede, em geral, testar normalidade.) [!h] [!h]

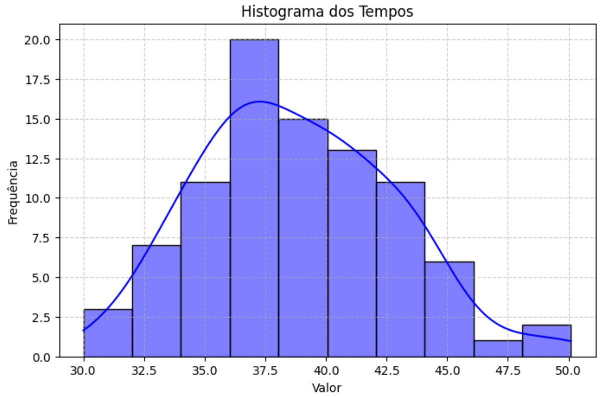


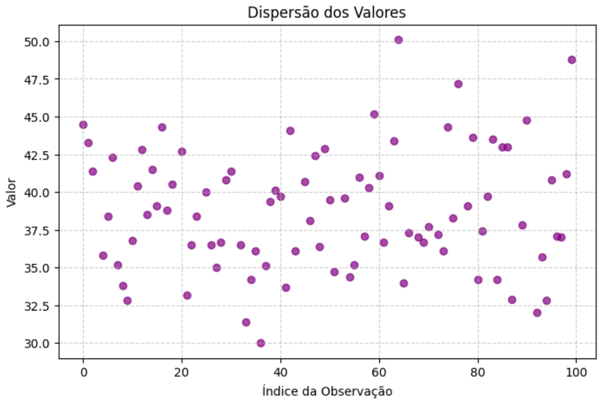
2) Proposição de Distribuições Candidatas

(A forma aproximadamente simétrica, média ≈ mediana e caudas moderadas sugerem uma Normal; como alternativa com suporte positivo e leve assimetria, pode-se considerar Lognormal. A seleção final utiliza aderência empírica.)

$$\text{Candidata A: } X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{com } \hat{\mu} = \bar{x},\ \hat{\sigma} = s.$$

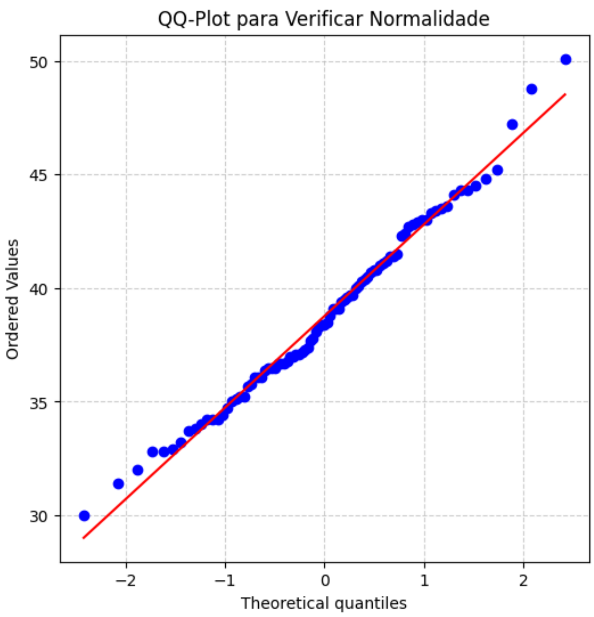
$$\text{Candidata B: } X \sim \text{Lognormal}(\mu_L, \sigma_L^2) \quad \text{com } \hat{\mu}_L = \overline{\ln X},\ \hat{\sigma}_L = s_{\ln X}.$$





3) Verificação Gráfica (diagnóstico)

(Histograma com curva Normal ajustada e Q–Q plot Normal: linhas quase retas e ausência de curvatura sistemática sustentam a hipótese Normal; leve cauda à esquerda é pequena.)



4) Teste de Aderência de Kolmogorov–Smirnov (KS)

(O KS compara a função distribuição empírica $F_n(x)$ com a teórica $F_0(x)$; a estatística é a maior distância vertical, e a decisão usa um limite crítico $\approx c(\alpha)/\sqrt{n}$.)

$$D \;=\; \sup_x |F_n(x) - F_0(x)|\,, \qquad \text{critério: } D \le \frac{c(\alpha)}{\sqrt{n}} \Rightarrow \text{não rejeita aderência.}$$

KS para Normal($\hat{\mu} = \bar{x}$, $\hat{\sigma} = s$) *(Estima-se μ e σ pelos moment estimators padrão; calcula-se $F_0(x)$ Normal e $F_n(x)$ empírico; D é o supremo das diferenças absolutas.)*

$$\hat{\mu} = 37{,}70, \quad \hat{\sigma} = 4{,}897, \qquad D_{\text{Normal}} \approx 0{,}052.$$

(Com $n = 100$ e $\alpha = 5\%$, usa-se $c(\alpha) \approx 1{,}36$, logo o ponto de corte é $1{,}36/\sqrt{100} = 0{,}136$. Como $0{,}052 < 0{,}136$, não há evidência para rejeitar a Normal.)

KS para Lognormal($\hat{\mu}_L, \hat{\sigma}_L$) *(Ajusta-se no domínio log, calcula-se a CDF lognormal correspondente e obtém-se a distância.)*

$$D_{\text{Lognormal}} \approx 0{,}078.$$

(Também menor que $0{,}136$, isto é, a Lognormal passa no KS; contudo, o ajuste Normal apresenta menor distância e, portanto, melhor aderência entre as duas candidatas.)

Conclusão

(Dado o formato quase simétrico e as métricas do KS, a distribuição Normal com parâmetros $\hat{\mu} = 37{,}70$ e $\hat{\sigma} = 4{,}897$ é uma modelagem adequada para os dados; a Lognormal também é aceitável, mas com ajuste ligeiramente inferior.)

$$\boxed{X \sim \mathcal{N}\big(37{,}70,\ 4{,}897^2\big) \text{ é uma escolha apropriada segundo EDA + KS.}}$$

3 Utilizando a mesma tabela, aplique os métodos Bootstrap e Bootstrap semi-paramétrico (utilizando os parâmetros obtidos de acordo com o teste de aderência da questão anterior) para determinar um intervalo de confiança para os parâmetros da distribuição que representa os dados com um nível de confiança de 95%.

1) Contexto e Objetivo

Dada a distribuição $\mathcal{N}(\mu \approx 37{,}70,\ \sigma \approx 4{,}897)$ identificada anteriormente como candidata adequada para modelar os dados, busca-se estimar intervalos de confiança (IC) de 95% para μ e σ utilizando dois métodos de reamostragem:

- **Bootstrap Não-Paramétrico**: reamostragem direta dos dados originais.
- **Bootstrap Semi-Paramétrico**: simulação de dados a partir da distribuição normal ajustada.

Este procedimento permite avaliar a variabilidade das estimativas sem depender estritamente de suposições analíticas de distribuição para os estimadores.

2) Método Bootstrap Não-Paramétrico

Descrição:

1. Gerar $B = 10.000$ amostras bootstrap de tamanho $n = 100$, com reposição, a partir dos dados originais.

2. Para cada amostra b, calcular:

$$\mu_b^* = \text{média amostral}, \quad \sigma_b^* = \text{desvio padrão amostral}.$$

3. Ordenar as B estimativas de μ^* e σ^*.

4. Determinar os ICs como os percentis 2,5% e 97,5% dessas distribuições.

Resultados:

$$\text{IC}_{95\%}(\mu) = [36{,}746;\ 38{,}636], \quad \text{IC}_{95\%}(\sigma) = [4{,}178;\ 5{,}557].$$

3) Método Bootstrap Semi-Paramétrico

Descrição:

1. Gerar $B = 10.000$ amostras de tamanho $n = 100$ a partir de $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, onde $\hat{\mu} = 37{,}70$ e $\hat{\sigma} = 4{,}897$.

2. Para cada amostra b, calcular μ_b^* e σ_b^* como antes.

3. Determinar os ICs como os percentis 2,5% e 97,5%.

Resultados:

$$\text{IC}_{95\%}(\mu) = [36{,}743;\ 38{,}644], \quad \text{IC}_{95\%}(\sigma) = [4{,}206;\ 5{,}553].$$

4) Comparação e Interpretação

Os intervalos obtidos pelos dois métodos são extremamente próximos, evidenciando que:

- A suposição de normalidade é compatível com a distribuição empírica.
- O método semi-paramétrico é mais eficiente quando a distribuição teórica é conhecida, mas o não-paramétrico confirma a robustez das estimativas sem depender desta suposição.

Conclusão: Com 95% de confiança, a média populacional está entre 36,74 e 38,64, e o desvio padrão entre 4,18 e 5,56. Ambos os métodos validam a modelagem normal para os dados.

$$\boxed{\text{Normalidade confirmada e estimativas consistentes entre os métodos.}}$$

4 Uma empresa de tecnologia deseja prever o salário dos funcionários com base em seus anos de experiência. Para isso, foi coletado um conjunto de dados contendo a experiência (em anos) e o salário de 10 funcionários.

Anos de Experiência	Salário
1	3500
2	3750
3	3900
4	4500
5	5100
6	5450
7	6000
8	6600
9	7000
10	8000

a. Determine a equação da reta de regressão linear y=ax+b, onde y representa o salário e x representa os anos de experiência.

b. Com base no modelo de regressão linear encontrado, estime o salário esperado para um funcionário com 12 anos de experiência.

1) Estatísticas descritivas essenciais

(A regressão linear do tipo $y = ax + b$ usa a inclinação a e o intercepto b, que são obtidos a partir das médias e dos somatórios de covariância e variância. As médias centralizam os dados para que os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ meçam variações em torno do centro.)

$$\bar{x} = \frac{1 + 2 + \cdots + 10}{10} = \frac{55}{10} = 5{,}5, \qquad \bar{y} = \frac{3500 + 3750 + \cdots + 8000}{10} = \frac{54\,800}{10} = 5\,480.$$

2) Somatórios fundamentais: S_{xx} e S_{xy}

(A inclinação a é a razão entre a covariância S_{xy} — variação conjunta de x e y — e a variância de x, S_{xx}. Logo, é necessário calcular os desvios em relação às médias, seus produtos e quadrados.)

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	3500	-4,5	-1 980	8 910	20,25
2	3750	-3,5	-1 730	6 055	12,25
3	3900	-2,5	-1 580	3 950	6,25
4	4500	-1,5	-980	1 470	2,25
5	5100	-0,5	-380	190	0,25
6	5450	0,5	-30	-15	0,25
7	6000	1,5	520	780	2,25
8	6600	2,5	1 120	2 800	6,25
9	7000	3,5	1 520	5 320	12,25
10	8000	4,5	2 520	11 340	20,25

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 8\,910 + 6\,055 + 3\,950 + 1\,470 + 190 - 15 + 780 + 2\,800 + 5\,320 + 11\,340 = 41\,800.$$

$$\sum (x_i - \bar{x})^2 = 20{,}25 + 12{,}25 + 6{,}25 + 2{,}25 + 0{,}25 + 0{,}25 + 2{,}25 + 6{,}25 + 12{,}25 + 20{,}25 = 82{,}5.$$

$$\Rightarrow \quad S_{xy} = 41\,800, \qquad S_{xx} = 82{,}5.$$

3) Coeficientes da reta: inclinação a e intercepto b

(A inclinação a mede a variação média do salário para cada ano adicional de experiência — razão direta entre a covariância S_{xy} e a variância S_{xx}. O intercepto b é o valor de y quando $x = 0$, ajustado pela média via $b = \bar{y} - a\bar{x}$.)

$$a = \frac{S_{xy}}{S_{xx}} = \frac{41\,800}{82{,}5} \approx 506{,}060606\ldots$$

$$b = \bar{y} - a\,\bar{x} = 5\,480 - (506{,}060606\ldots \times 5{,}5) = 5\,480 - 2\,783{,}333333\ldots = 2\,696{,}666666\ldots$$

$$\Rightarrow \qquad \boxed{\hat{y} = 506{,}\overline{06}\,x + 2\,696{,}\overline{66}}$$

4) Previsão para 12 anos de experiência

(A previsão pontual aplica o modelo ajustado ao novo x; interpreta-se como o salário esperado para 12 anos de experiência, dadas as tendências lineares extraídas dos dados.)

$$\hat{y}(12) = 506{,}\overline{06} \times 12 + 2\,696{,}\overline{66} = 6\,072{,}\overline{72} + 2\,696{,}\overline{66} = 8\,769{,}\overline{38}.$$

Resposta

(a) **Equação da reta de regressão:** $\hat{y} = 506{,}\overline{06}\,x + 2\,696{,}\overline{66}$.

(b) **Salário estimado para 12 anos:** $\hat{y} = 8\,769{,}\overline{38}$ (aprox. R$ 8.769,39).