ChatGPT como Ferramenta de Geração de Dados para Modelagem Estatística: Uma Abordagem Prática na Disciplina de Data Science

Italo Pinto Rodrigues[1,2]; 0000-0002-7558-4958
Alexandra Matos Campos[1]; 0009-0008-3250-1252

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
2 – INPE, Instituto Nacional de Pesquisas Espaciais, SP.
italoprodrigues@gmai.com

Resumo: O ensino de estatística, particularmente no contexto da disciplina de Data Science e Inteligência de Mercado, frequentemente enfrenta o desafio de manter os estudantes engajados e tornar o aprendizado atrativo. Este artigo aborda essa questão, explorando o uso do ChatGPT como uma ferramenta inovadora para a geração de dados personalizados para modelagem estatística. Ao interagir com o ChatGPT, foi possível produzir conjuntos de dados seguindo a distribuição de Weibull, proporcionando aos estudantes uma experiência mais realista e contextualizada. Os resultados obtidos confirmam a eficácia do ChatGPT nesse papel, com uma precisão notável na modelagem, destacando seu potencial como uma ferramenta pedagógica revolucionária.

Palavras-chave: ChatGPT. Data Science. Distribuição de Weibull. Engenharia de Produção.

INTRODUÇÃO

A estatística fornece as ferramentas necessárias para entender a variabilidade e a incerteza, presentes nos dados coletados. Portanto, o ensino eficaz de estatística é crucial para preparar os futuros profissionais de dados para os desafios da profissão (SHAH, 2023). Nesse sentindo, a estatística, é um conteúdo fundamental na disciplina Data Science e Inteligência de Mercado do Centro Universitário de Volta Redonda (UniFOA). A capacidade de coletar, analisar e interpretar dados é essencial para a tomada de decisões informadas em um mundo cada vez mais orientado por dados (NUNES; SANTOS; ROCHA, 2023).

Tradicionalmente, o ensino deste conteúdo em engenharia era conduzido com foco no docente. No entanto, essa abordagem clássica tem dado lugar a métodos que buscam enriquecer a assimilação de conhecimentos, incentivando o engajamento direto dos alunos no processo educativo (PRINCE; FELDER, 2006). Nesse sentido, as metodologias ativas de ensino estão reformulando o cenário educacional, incentivando uma aprendizagem mais engajada e dinâmica. Diversas estratégias, como aprendizagem orientada a projetos, gamificação e sala de aula invertida, estão sendo

adotadas para ampliar a motivação e compreensão dos alunos (GOMEZ-DEL RIO; RODRIGUEZ, 2022).

Dentro do contexto das metodologias ativas de ensino, a ascensão da Inteligência Artificial (IA) e dos modelos de linguagem representa uma confluência entre pedagogia e tecnologia avançada. O ChatGPT, como um produto dessa intersecção, capitaliza os avanços da IA para potencializar a entrega educacional, estabelecendo um novo paradigma na integração de tecnologia e práticas pedagógicas inovadoras. (BROWN et al., 2020; JAVAID et al., 2023).

Com isso, percebeu-se a do ChatGPT de gerar dados customizados, o que pode ser particularmente útil para simulações estatísticas e experimentos pedagógicos. A possibilidade de criar conjuntos de dados específicos e adaptados às necessidades do educador ou pesquisador abre portas para uma abordagem mais prática e contextualizada do ensino de estatística.

O objetivo principal deste artigo é explorar o potencial do ChatGPT como ferramenta de geração de dados para modelagem estatística. Desse modo, será possível entender como essa integração pode enriquecer o processo de aprendizagem, proporcionando experiências educacionais mais dinâmicas e adaptadas às demandas contemporâneas.

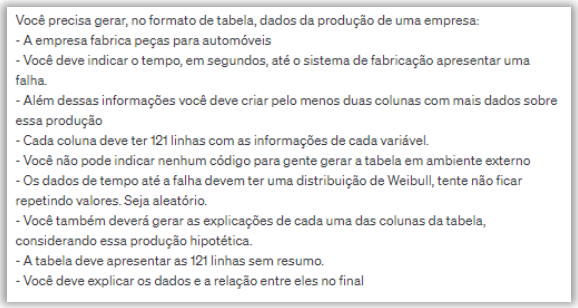
METODOLOGIA

O experimento seguiu três etapas: interação com ChatGPT, preparação dos dados e modelagem estatística. Presumindo que o usuário esteja registrado no ChatGPT 3.5, o objetivo principal da interação é adquirir dados para modelagem estatística relacionados ao domínio da Engenharia de Produção, conforme a matriz curricular da disciplina.

Interação com ChatGPT

Para iniciar a interação é necessário criar um prompt, isto é, uma conversa com o ChatGPT. O ideal é que nesta etapa sejam passadas informações detalhadas. Conforme a Figura 1, detalhes específicos foram fornecidos ao ChatGPT. No entanto, nem todas as instruções, como "A tabela deve apresentar as 121 linhas sem resumo", foram completamente atendidas, como visto na Figura 2. Se o ChatGPT não atender às expectativas, é essencial refinar a interação até alcançar o resultado desejado.

Figura 1 – Primeira interação com o ChatGPT, com informações detalhadas.



Fonte: Os Autores (2023).

Deve-se ressaltar que o ajuste fino dependerá do objetivo esperado. Para exemplificar, neste experimento foram indicados dois ajustes, conforme Figura 4.

O primeiro ajuste foi necessário pois o ChatGPT criou os dados com uma relação proporcional, então para tornar mais verossímil, foi dado um comando para os dados fossem gerados segundo uma distribuição. Além disso, a ideia neste experimento é fazer com que o estudante modele os dados estatisticamente. Então, gerar os dados conforme uma distribuição auxilia no processo de análise dos resultados. Caso o professor queira omitir essa informação, isso pode ser feito. Neste experimento, decidiu-se por fazer a modelagem cooperativamente.

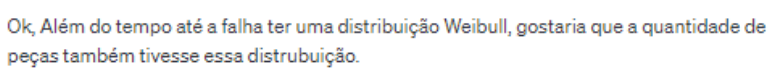
Já o segundo ajuste solicitado, se deu por conta da tabela apresentada na Figura 2. Sendo assim, o experimento poderia ser realizado com 121 amostras.

Figura 2 – Tabela com 6 amostras em vez de 121.

Nº	Tempo até Falha (s)	Quantidade Produzida	Qualidade da Peça
1	152	76	Alta
2	98	49	Média
3	210	105	Alta
4	67	33	Baixa
5	134	67	Alta
...	...	...	...
121	176	88	Média

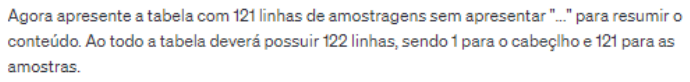
Fonte: Os Autores (2023).

Figura 3 – Prompt para melhorar a distribuição dos dados.



Fonte: Os Autores (2023).

Figura 4 – Prompt para melhorar a visualização dos dados.



Fonte: Os Autores (2023).

Preparação dos dados

Ao utilizar o ChatGPT para gerar dados destinados à modelagem em softwares como Excel ou Planilhas Google, pode ser essencial tratar a tabela produzida. Isso pode envolver ajustes como a modificação do separador decimal de ponto para vírgula e a concatenação de múltiplas tabelas geradas pela ferramenta para alcançar o número desejado de amostras.

Modelagem estatística

A modelagem estatística foi realizada considerando os dados em rol, isto é, ordenados do menor para o maior e foram realizadas as análises apresentadas na Tabela 1.

Tabela 1 – Modelagem estatística.

Item	Descrição	Equação
Média aritmética simples	A média frequentemente utilizada para resumir e representar de forma concisa um conjunto de dados.	1
Mediana	É uma medida de tendência central que descreve o valor que separa a metade superior da metade inferior de um conjunto de dados.	Se a quantidade de amostras for ímpar, Equação 2, se não, Equação e 3
Moda	A moda representa o valor mais recorrente em um conjunto de dados. Caso não haja repetições no conjunto, significa que ele não tem uma moda definida.	-

Fonte: (RAMACHANDRAN; TSOKOS, 2021).

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K} x_i \quad (1)$$

Onde, x_i são as amostras e K é a quantidade total de amostras

$$R_M = \frac{N+1}{2} \quad (2)$$

Onde R_M representa a posição do valor da mediana.

$$M = \frac{d_{\frac{N}{2}} + d_{\left(\frac{N}{2}\right)+1}}{2} \quad (3)$$

Onde M é a mediana, calculada pela média dos valores centrais, em que d representa o valor nas posições $\frac{N}{2}$ e $\left(\frac{N}{2}\right)+1$.

Além disso, será montada uma tabela de frequência para geração do histograma e, posteriormente modelagem da distribuição, conforme Tabela 2.

Tabela 2 – Tabela de Frequência.

Item	Descrição	Equação
Amplitude	Representa a variação entre o valor máximo e o mínimo observados.	4
Frequência absoluta	Refere-se à frequência com que um elemento se manifesta na amostra.	5
Frequência relativa		6
Frequência absoluta acumulada	Representa a acumulação da frequência absoluta do valor da variável i com todas as frequências que a precedem.	N_i
Frequência relativa acumulada		7
Classes	As classes servem como uma ferramenta para agrupar e simplificar a variedade de elementos presentes em uma amostra.	8
Intervalo de Classe		9

Fonte: (RAMACHANDRAN; TSOKOS, 2021).

$$H = Valor_{MÁXIMO} - Valor_{MÍNIMO} \quad (4)$$

$$n = \sum_{1}^{K} n_i \quad (5)$$

Onde n_i é a frequência absoluta, n número total de dados na amostra.

$$f_i = \frac{n_i}{n} \quad (6)$$

$$F_i = \frac{N_i}{n} \quad (7)$$

$$k \cong \sqrt{K} \quad (8)$$

$$h \cong \frac{H}{k} \quad (9)$$

Para a modelagem da distribuição Weibull foi utilizada a Equação 10.

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} * e^{-\left(\frac{t}{\eta}\right)^{\beta}} \quad (10)$$

Onde, β é o parâmetro de forma (ou inclinação) e η é o tempo de vida.

RESULTADOS E DISCUSSÃO

Após as diversas perguntas, o ChatGPT gerou as 121 amostras. A tabela foi gerada com as seguintes colunas:

- Tempo até a Falha (em segundos): Indica o tempo, em segundos, até o sistema de fabricação apresentar uma falha. Os dados são gerados com base em uma distribuição de Weibull, que é comumente usada para modelar tempos de vida de produtos.
- Quantidade Produzida: Esta coluna representa a quantidade de peças produzidas antes da falha ocorrer. A ideia é que quanto mais tempo o sistema funciona sem falhas, mais peças são produzidas.

A Tabela 3 apresenta o resultado dos itens elencados na Tabela 1. A Tabela 4 apresenta os resultados dos itens elencados na Tabela 2, e, a partir destes dados, foi construído o histograma, conforme Figura 5. O erro médio quadrático (RMS) demonstra que a distribuição Weibull realmente se aproxima dos dados, conforme

Figura 6.

Tabela 3 – Resultados da modelagem estatística.

Item	Resultado
Média aritmética simples	$\cong 158{,}76$ segundos
Mediana	160 segundos
Moda	160 segundos

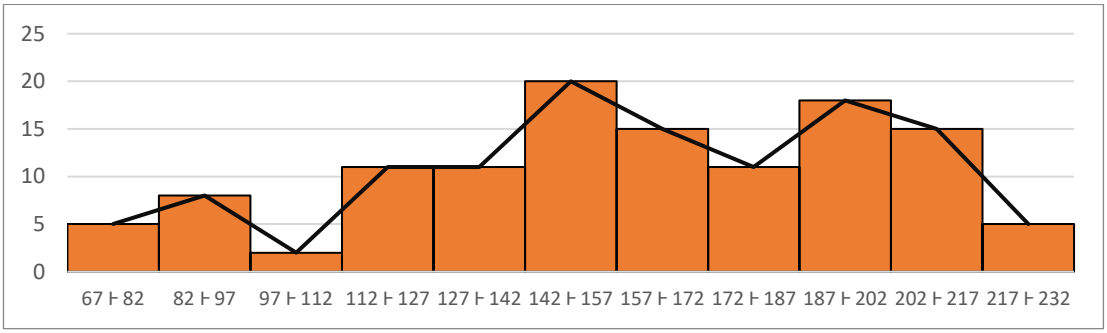
Fonte: Os Autores (2023).

Tabela 4 – Modelagem estatística.

i	Classes	n_i	Ponto Médio	N_i	f_i	F_i	$f(t)$	Quadrado do Desvio (f_i e $f(t)$)
1	67 ⊢ 82	5	74,5	5	0,041	0,041	0,00115649	0,00000255465
2	82 ⊢ 97	8	89,5	13	0,066	0,107	0,00219922589	0,00000487742
3	97 ⊢ 112	2	104,5	15	0,017	0,124	0,00369058802	0,00000670116
4	112 ⊢ 127	11	119,5	26	0,091	0,215	0,00556058945	0,00000025002
5	127 ⊢ 142	11	134,5	37	0,091	0,306	0,00755599317	0,00000223618
6	142 ⊢ 157	20	149,5	57	0,165	0,471	0,00921991134	0,00000323774
7	157 ⊢ 172	15	164,5	72	0,124	0,595	0,00999004175	0,00000297762
8	172 ⊢ 187	11	179,5	83	0,091	0,686	0,00944953906	0,00001148487
9	187 ⊢ 202	18	194,5	101	0,149	0,835	0,00763014452	0,00000523133
10	202 ⊢ 217	15	209,5	116	0,124	0,959	0,00511675285	0,00000990808
11	217 ⊢ 232	5	224,5	121	0,041	1,000	0,00275854559	0,00000000001
					1		RMS	0,002120443

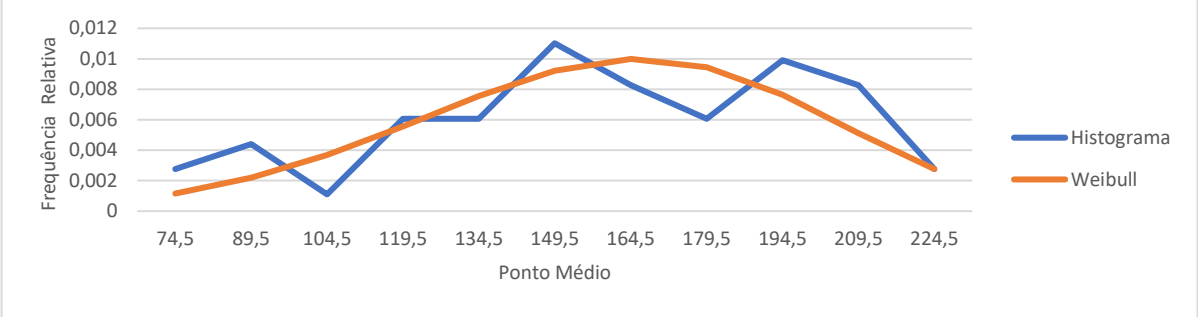
Fonte: Os Autores (2023).

Figura 5 – Histograma dos dados.



Fonte: Os Autores (2023).

Figura 6 – Comparação entre os dados gerados pelo ChatGPT e a Distribuição Weibull.



Fonte: Os Autores (2023).

CONCLUSÕES

O ChatGPT provou ser uma ferramenta útil para a geração de dados no contexto educacional, permitindo a contextualização de exemplos durante as aulas. Sua capacidade de fornecer conjuntos de dados distintos para cada estudante é particularmente notável, especialmente considerando que os dados gerados, ao seguir a distribuição Weibull, apresentaram um erro mínimo na modelagem, evidenciando sua precisão e aplicabilidade.

Adicionalmente, os dados produzidos por esta ferramenta oferecem uma oportunidade valiosa para os alunos exercitarem a modelagem estatística básica. Através da análise desses conjuntos de dados, os estudantes podem praticar conceitos fundamentais, como média, moda e mediana, solidificando sua compreensão e habilidade em estatística aplicada e Data Science.

REFERÊNCIAS

BROWN, Tom B. et al. Language Models are Few-Shot Learners. *[S. l.]*, 2020. Disponível em: http://arxiv.org/abs/2005.14165.

GOMEZ-DEL RIO, T.; RODRIGUEZ, J. Design and assessment of a project-based learning in a laboratory for integrating knowledge and improving engineering design skills. **Education for Chemical Engineers**, *[S. l.]*, v. 40, p. 17–28, 2022. DOI: 10.1016/j.ece.2022.04.002.

JAVAID, Mohd; HALEEM, Abid; SINGH, Ravi Pratap; KHAN, Shahbaz; KHAN, Ibrahim Haleem. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. **BenchCouncil Transactions on Benchmarks, Standards and Evaluations**, *[S. l.]*, v. 3, n. 2, p. 100115, 2023. DOI: 10.1016/j.tbench.2023.100115.

NUNES, P.; SANTOS, J.; ROCHA, E. Challenges in predictive maintenance – A review. **CIRP Journal of Manufacturing Science and Technology**, *[S. l.]*, v. 40, p. 53–67, 2023. DOI: 10.1016/j.cirpj.2022.11.004. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1755581722001742.

PRINCE, Michael J.; FELDER, Richard M. Inductive Teaching and Learning Methods: Definitions, Comparisons, and Research Bases. **Journal of Engineering Education**, *[S. l.]*, v. 95, n. 2, p. 123–138, 2006. DOI: 10.1002/j.2168-9830.2006.tb00884.x. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2006.tb00884.x.

RAMACHANDRAN, Kandethody M.; TSOKOS, Chris P. Descriptive statistics. *Em*: **Mathematical Statistics with Applications in R**. [s.l.] : Elsevier, 2021. p. 1–40. DOI: 10.1016/B978-0-12-817815-7.00001-4.

SHAH, Chirag. The past, the present, and the future of information and data sciences: A pragmatic view. **Data and Information Management**, *[S. l.]*, v. 7, n. 1, p. 100028, 2023. DOI: 10.1016/j.dim.2023.100028. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S2543925123000025.