Estimação da Média em Amostras Obtidas por Diferentes Técnicas de Amostragem

Breno Cauã Rodrigues da Silva

2024-10-09

Resumo

Este relatório ...

Palavras-Chave: Amostragem;

Abstract

This report ...

Keywords: Sampling;

1 Introdução

A estimativa precisa de parâmetros populacionais, como a média (μ), é fundamental em diversas áreas do conhecimento, especialmente nas ciências estatísticas. A obtenção de uma boa estimativa depende diretamente da técnica de amostragem utilizada, uma vez que diferentes métodos podem influenciar a representatividade dos dados em relação à população. Técnicas de amostragem bem escolhidas garantem que os resultados de uma amostra sejam generalizáveis e reduzam possíveis vieses no processo de inferência.

Este projeto tem como objetivo comparar diferentes técnicas de amostragem na estimativa do parâmetro μ (média populacional), investigando como cada método afeta a precisão e a eficiência da estimativa. Técnicas como amostragem aleatória simples, estratificada, sistemática e por conglomerados serão abordadas, utilizando conjuntos de dados reais e simulados. Ao analisar as diferenças nos resultados, será possível determinar qual técnica fornece a melhor estimativa da média em diferentes cenários populacionais.

Além disso, este estudo visa explorar as vantagens e limitações de cada técnica, destacando a importância da escolha do método adequado em pesquisas empíricas. A análise será conduzida utilizando métodos estatísticos apropriados, como a construção de intervalos de confiança e testes de hipóteses para a média, possibilitando uma avaliação rigorosa da performance de cada técnica.

2 Materias e Métodos

2.1 Conjunto de Dados

Para a realização deste estudo, utilizamos um conjunto de dados que contém informações de beneficiários de seguros de saúde nos Estados Unidos. O conjunto de dados foi originalmente disponibilizado em domínio público, com o objetivo de explorar a relação entre diversas variáveis e os custos médicos individuais cobrados pelo seguro. Tal conujunto possui 2.772 observações e 7 variáveis, e pode ser obtido diretamente no **Kaggle**. O Quadro (2.1), mostra as variáveis do conjunto de dados e suas descrições.

Quadro 2.1: Descrição das Variáveis do Conjunto de Dados em Análise.

Variável	Descrição
age	Idade do beneficiário primário
sex	Gênero do beneficiário (masculino ou feminino)
bmi	Índice de Massa Corporal, uma medida objetiva que relaciona altura e peso corporal. Esse índice é amplamente utilizado para categorizar indivíduos em faixas de peso ideal, abaixo ou acima do ideal
smoker	Status de fumante do beneficiário (sim ou não)
region	Local de residência do beneficiário, dividido em quatro regiões: nordeste, sudeste, noroeste e sudoeste dos Estados Unidos
charges	Custos individuais cobrados pelo seguro de saúde, representando o valor monetário dos serviços médicos utilizados

Fonte: Elaborado pelo autor.

Este conjunto de dados foi escolhido pela diversidade de suas variáveis, o que possibilita a aplicação de diferentes técnicas de amostragem, considerando tanto variáveis numéricas quanto categóricas. A utilização de uma variável de resposta contínua, como o custo dos serviços médicos (charges), proporciona uma oportunidade de estudar como a estimativa do parâmetro μ (média dos custos) é impactada pelas diferentes estratégias de amostragem.

2.2 Software Utilizado

Para conduzir as análises e estimativas neste estudo, foi utilizada a linguagem de programação Python, empregando a IDE **Google Colaboratory**. As seguintes bibliotecas foram utilizadas nas diversas etapas do processo:

- **Numpy:** Para operações matemáticas, lógicas e estatísticas eficientes em arrays multidimensionais ou matrizes (Harris et al. 2020);
- **Pandas:** Para manipulação e análise de dados, oferecendo estruturas de dados flexíveis e poderosas (McKinney 2010);
- **Matplotlib:** Para criação de visualizações gráficas (Hunter 2007);
- **Seaborn:** Complementar ao Matplotlib, oferece uma interface de alto nível para criação de gráficos estatísticos atrativos e informativos (Waskom 2021).
- **Scipy:** Para computação científica, em específico Testes de Hipóteses e Intervalos de confiança através da sub-biblioteca `scipy.stats` (Virtanen et al. 2020).

Escrita, estrutura e editoração científica deste relatório foram feitas por meio do ***R Markdown***.

2.3 Metodologia

O desenvolvimento de estimação do parâmetro μ, seguiu-se os passos abaixo:

1. **Processo de Amostragem**: A depender do método adotado, será abordado nesse trabalho as seguintes técnicas: *Amostragem Aleatória Simples (AAS)*, *Amostragem Aleatória Estratificada (AAE)*, *Amostragem Sistemática (AS)*, *Amostragem Sistemática Estratificada (ASE)*. Todos os métodos foram sem reposição. Será falado mais sobre estes métodos a decorrer do trabalho.
2. **Estatística Descritiva:** Esta etapa envolveu análises estatísticas através de visualizações gráficas para compreender a distribuição, variabilidade, padrões e relações nos dados.
3. **Estatística Inferencial:** Estimação, Teste de Hipóteses e Intervalos de Confiaça.
4. **Avaliação dos Métodos de Amostragem:** Comparação de desempenho das *Técnicas de Amostragem* utilizadas.

Para Inferências realizadas nesse trabalho, o nível de significância foi fixado em 5%, logo, $\alpha = 0,05$.

O tamanho da amostra foi calculado usando a fórmula simplificada

$$n = \frac{N \times n_0}{N + n_0},$$

onde $n_0 = \frac{1}{\varepsilon^2}$, e ε representa a margem de erro amostral admitida. Fixando $\varepsilon = 0,05$, ou seja, uma margem de erro de 5%, obtemos $n = 350$.

3 Referencial Teórico

A referência usada para tais definições quanto a ideia e essência de cada método de amostragem podem ser vistas com mais afinco em Bussab and BOLFARINE (2005), e para estudos ainda mais elaborados deixa-se a referência de Cochran (1977).

3.1 Amostragem Aleatória Simples (AAS)

Em um processo de AAS, a ideia é bem ituitiva e por isso, redudantemente, acaba por se tornar mais simples. Parte-se do príncipio que tenha-se uma lista enumerada de $1, 2, \cdots, N$ unidades populacionais, são selecionadas, aleatóriamente, n unidades com probabilidades iguais.

3.2 Amostragem Aleatória Estratificada (AAE)

Em um processo de AAE, a ideia muda em alguns aspectos em comparação a AAS. O foco da AAE é separar a população em subgrupos, chamados de estratos, de acordo com determinada característica. Basicamente,

$$N = N_1 + N_2 + \cdots + N_H,$$

onde H é o número de estratos (divisões) feitas na população.

Com um objetivo de tornar grupos heterogêneos em grupos mais homôgenea, melhorando a estimativa e contralando sua variabilidade. Feita essa divisão, realiza-se uma AAS em cada estrato formado. Desta forma, monta-se o conjunto de dados amostrais :

$$n = n_1 + n_2 + \cdots + n_h.$$

3.3 Amostragem Sistemática (AS)

Seja uma população de N unidades, onde é possível se chegar a relação $N = kn$, onde k é um número inteiro.

A amostra obtida através de um processo AS, tem com primeiro elemento um número sorteado aleatoriamente, seguindo o raciocínio da AAS, entre 1 e k, sendo usado um sistema de referência em que a população está enumerada de $1, 2, \cdots, N$. Desta forma, as demais unidades ou elementos que irão compor a amostra, são obtidos em intervalos de espaços iguais a k.

3.4 Amostragem Sistemática Estratificada (ASE)

Como já se sabe os conceitos de AAE e AS . Acaba que por consequência, a dedução do entendimento da ideia por da AS é simplemente, dividir a população em estratos conforme determinada característica e, dentro e cada estrato, aplicar uma AS.

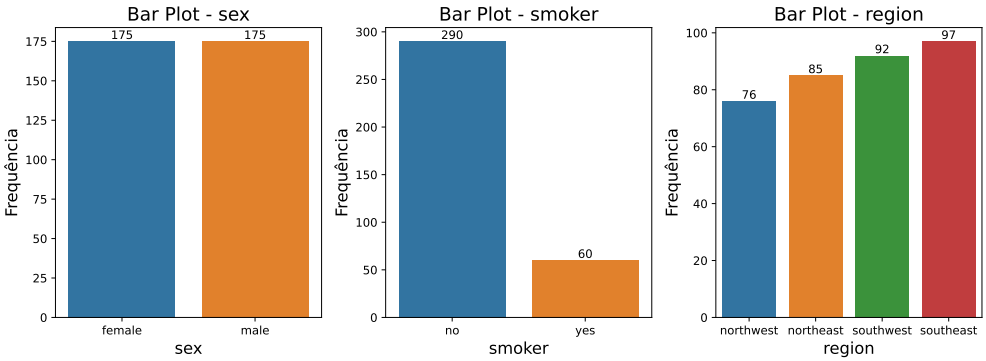
4 Resultados e Discrusão

Para cada técnica de amostragem, foi feito o processo descrito na Metodologia deste trabalho. Veja a seguir, os resultados para AAS.

4.1 Resultados para Amostragem Aleatória Simples

4.1.1 Estatística Descritiva

Figura 4.1: Gráfico de Barras para as variáveis qualitativas do Conjunto de Dados.



Fonte: Elaborado pelo autor.

A Figura (4.1), tem como objetivo averiguar as proporções de cada classe das variáveis qualitativas. Com o principal foco de já selecionar as variáveis que serão usadas para estratificação.

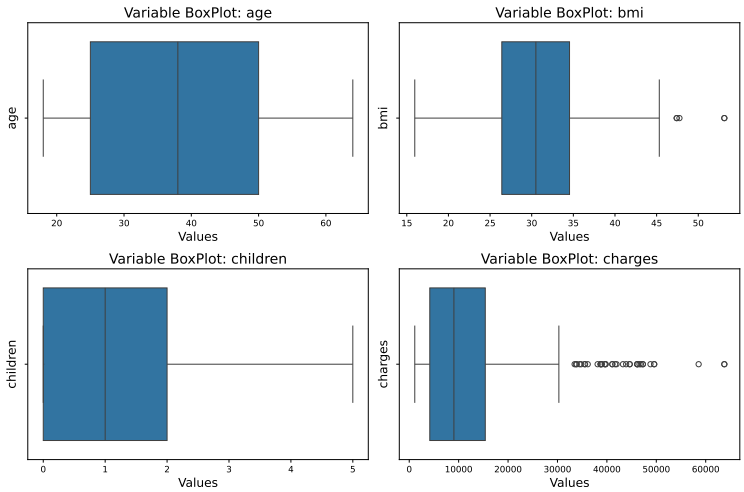
Como é possível notar, a variável *sex* é perfeita distribuída, tendo 50% de cada gênero, provavelmente, os dados devem seguir aproximadade essa distribuição de classes ou tal coincidência, por assim dizer, se deve ao processo de amostragem, sendo ele, aleatório.

Uma variável bem desbalanceada é a *smoker*, se é ou não fumante, tendo 82,86% aproximadamente de suas observações averiguadas como não fumante e, apenas 17,14% fumante, mostrando um grande desbalanceamento.

Já a variável *region*, talvez não pareça, mas as observações estão bem distribídas. Sem grandes diferênças, como por exemplo, a diferênça entre a proporção de pessoas da região noroeste (menor proporção) e sudeste (maior proporção) é de exatos 6%.

A seguir, veja a Figura (4.2), mostra o gráfico de boxplot das variáveis *age*, *bmi*, *children* e *charges*.

Figura 4.2: Boxplot das variáveis numéricas do Conjunto de Dados.



Fonte: Elaborado pelo autor.

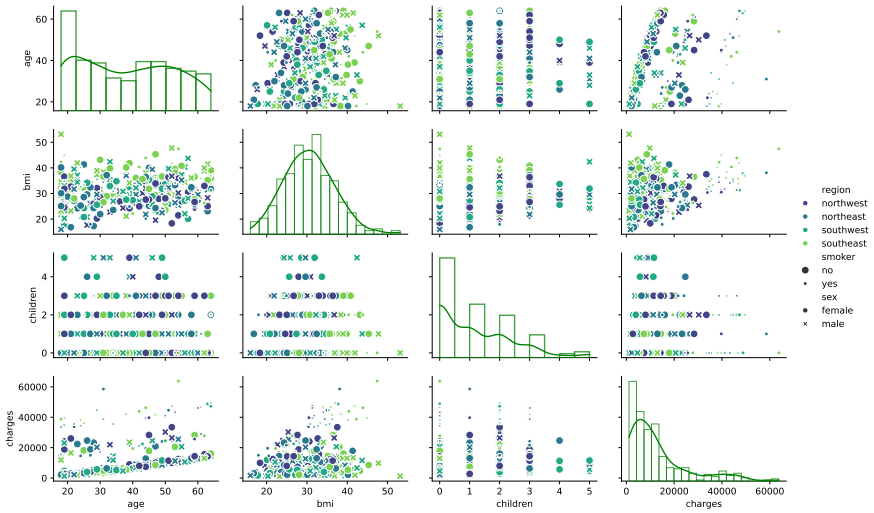
Ao observar a Figura (4.2), nota-se que apenas as variáveis *bmi* e *charges* apresentaram *outliers*, valores discrepantes ou muito distante do esperado. Apesar de apresentar *outliers*, a variábel *bmi*, que representa o *Índice de Massa Corporal*, aparenta apresentar uma distribuição mais simétrica em relação as outras, tal supeita pode ser averiguda na Figura (4.3), onde é apresentado os histogramas das variáveis numéricas.

A variável *children*, mostra uma grande concetração de observações dos pacientes que possuem de 0 a 2 filhos, começando a se tornar mais dispersos entre as observações com 3 ou mais. A variável *age*, aparenta estar mais próxima de uma uniforme, como pode-se ver no boxplot, seus quantis estão mais bem "esparçados".

Como já formam identificados os possíveis outiliers, vamos para um gráfico de combinação. Onde a diagonal principal da Figura a seguir, é contituida pelos histogramas das variáveis

numéricas, importante para averiguar as suspeitas quanto a distribuição das observações de cada variável. Veja a seguir, a Figura (4.3).

Figura 4.3: Gráfico Combição - Diagrama de Dispersão e Histograma com KDE - das variáveis numéricas do Conjunto de Dados.



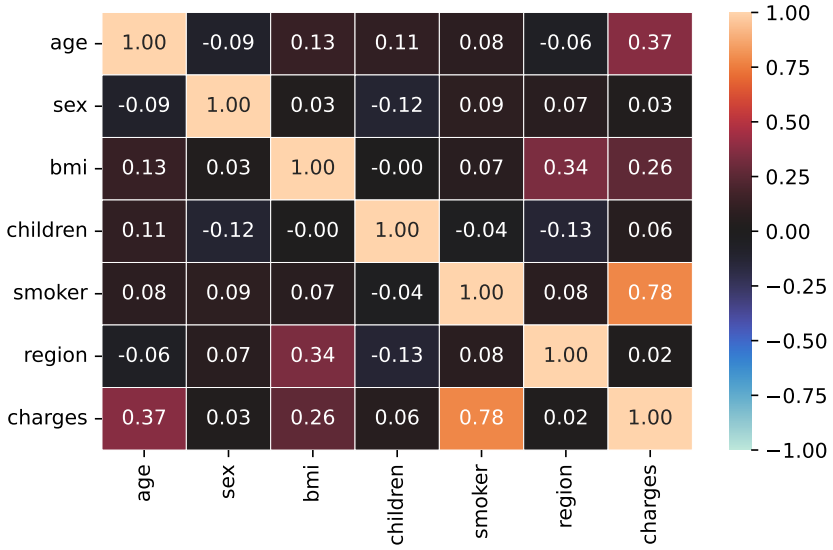
Fonte: Elaborado pelo autor.

Ao olhar para Figura (4.3) pode-se confirmar algumas suspeitas, como a da varável *age* esar próxima de uma uniformidade, apesar de ter consideráveis desvios como uma concentração maior à esquerda do histograma. Outra suspeita confirmada, foi de que a variável *bmi*, ser a mais próxima de uma distribuição simétrica, talvez até passase em um teste de normalidade, mas não é algo que será feito neste trabalho. Uma concentração maior no lado esquerdo , também pode ser vista, no histograma da variável *children*, já acusado no boxplot. Vale salientar que no histograma se torna notório a presença de outliers na variável *charges*. Tudo isso pode ser visto ao olhar para os histogramas de cada variável, localizados na diagonal da grade de gráficos.

Entretanto, o mais interessante da Figura (4.3) talvez não sejam os histogramas, mas sim os diagramas de dispersão. Que possibilitam analisar a relação entre as variáveis do conjunto de dados umas com as outras. E ainda, cada obervação está "classificada" de acordo com as classes de cada variável qualitativa que contém a base de dados. Tornando a visualização da observação ainda mais interessante. Por exemplo, no diagrama de dispersão em que se tem *charges vs bmi*, pode-se identificar algum tipo de relação positiva, porém, é curioso notar que homens tendem a pagar mais caro que mulheres, principalmente se for fumante. Ou ainda, pessoas que moram nas regiões sudoeste (*southwest*) e sudeste (*southeast*) tem maior índice massa corporal e com isso tendem a pagar mais caro.

Porém visualmente ainda é difil cravar se uma variável tem relação com outra, por isso foi calculada a *Matrix de Correlação Linear de Pearson*, para mensurar tais relações e se elas de fato existem. Veja a Figura (4.4).

Figura 4.4: Mapa de Calor da Matriz de Correlação de Pearson para o Conjunto de Dados.



Fonte: Elaborado pelo autor.

4.1.2 Estatística Inferêncial

4.1.2.1 Estimação Pontual

4.1.2.2 Estimação Intervalar - Intervalo de Confiança (IC)

4.1.2.3 Teste de Hipóteses

4.2 Resultados para Amostragem Aleatória Estratificada

Referências

Bussab, W de O, and Heleno BOLFARINE. 2005. "Elementos de Amostragem." *São Paulo: Edgar Blucher*.

Cochran, William G. 1977. "Sampling Techniques." *John Wiley& Sons*.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/ 10.25080/Majora-92bf1922-00a .

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for

Scientific Computing in Python." *Nature Methods* 17: 261–72. https://doi.org/10.1038/s41592-019-0686-2.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.