House of excellence

Consultores Responsáveis:
Estatiano 1
Estatiano 2
...
Estatiano n

Requerente:
ESTAT

Brasília, 12 de outubro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

O projeto tem como objetivo auxiliar João Neves, proprietário da academia de alta performance House of Excellence, na otimização do desempenho de seus atletas de elite, com base em análises estatísticas de suas participações nas edições dos Jogos Olímpicos de 2000 a 2016. O foco das análises é identificar padrões de desempenho, características físicas e fatores relacionados às conquistas de medalhas, oferecendo insights valiosos para melhorar a preparação e a performance futura dos atletas.

Segundo parágrafo: detalhar quais são as análises que serão feitas. TODOS OS TIPOS DE ANÁLISES, se é análise descritiva , teste de hipóteses, regressão. Descrever um pouco sobre o banco de dados(quantidade de variáveis, tipos de variáveis, etc)..

As análises foram realizadas utilizando o software R, versão 4.4.1, com pacotes especializados para manipulação de dados, visualização gráfica e modelagem estatística.

2 Referencial Teórico

2.1 Média

A média é a soma das observações dividida pelo número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Com:

- $i = 1, 2, ..., n$
- $n =$ número total de observações

2.2 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é o valor que deixa metade das observações abaixo dela e metade dos dados acima.

Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, \text{para n ímpar} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{para n par} \end{cases}$$

2.3 Quartis

Os quartis são separatrizes que dividem o conjunto de dados em quatro partes iguais. O primeiro quartil (ou inferior) delimita os 25% menores valores, o segundo representa a mediana, e o terceiro delimita os 25% maiores valores. Inicialmente deve-se calcular a posição do quartil:

- Posição do primeiro quartil P_1:

$$P_1 = \frac{n+1}{4}$$

- Posição da mediana (segundo quartil) P_2:

$$P_2 = \frac{n+1}{2}$$

- Posição do terceiro quartil P_3:

$$P_3 = \frac{3 \times (n+1)}{4}$$

Com n sendo o tamanho da amostra. Dessa forma, $X_{(P_i)}$ é o valor do i-ésimo quartil, onde $X_{(j)}$ representa a j-ésima observação dos dados ordenados.

Se o cálculo da posição resultar em uma fração, deve-se fazer a média entre o valor que está na posição do inteiro anterior e do seguinte ao da posição.

2.4 Variância

A variância é uma medida que avalia o quanto os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados.

2.4.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Com:

- $X_i = i$-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.4.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.5 Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Ele avalia o quanto os dados estão dispersos em relação à média.

2.5.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

Com:

- $X_i =$ i-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.5.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.6 Coeficiente de Variação

O coeficiente de variação fornece a dispersão dos dados em relação à média. Quanto menor for o seu valor, mais homogêneos serão os dados. O coeficiente de variação é considerado baixo (apontando um conjunto de dados homogêneo) quando for menor ou igual a 25%. Ele é dado pela fórmula:

$$C_V = \frac{S}{\bar{X}} \times 100$$

Com:

- $S =$ desvio padrão amostral
- $\bar{X} =$ média amostral

2.7 Coeficiente de Assimetria

O coeficiente de assimetria quantifica a simetria dos dados. Um valor positivo indica que os dados estão concentrados à esquerda em sua função de distribuição, enquanto um valor negativo indica maior concentração à direita. A fórmula é:

$$C_{Assimetria} = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S} \right)^3$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $S =$ desvio padrão amostral
- $n =$ tamanho da amostra

2.8 Curtose

O coeficiente de curtose quantifica o achatamento da função de distribuição em relação à distribuição Normal e é dado por:

$$Curtose = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S} \right)^4 - 3$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $S =$ desvio padrão amostral
- $n =$ tamanho da amostra

Uma distribuição é dita mesocúrtica quando possui curtose nula. Quando a curtose é positiva, a distribuição é leptocúrtica (mais afunilada e com pico). Valores negativos indicam uma distribuição platicúrtica (mais achatada).

2.9 Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos. A figura abaixo ilustra um exemplo de boxplot.

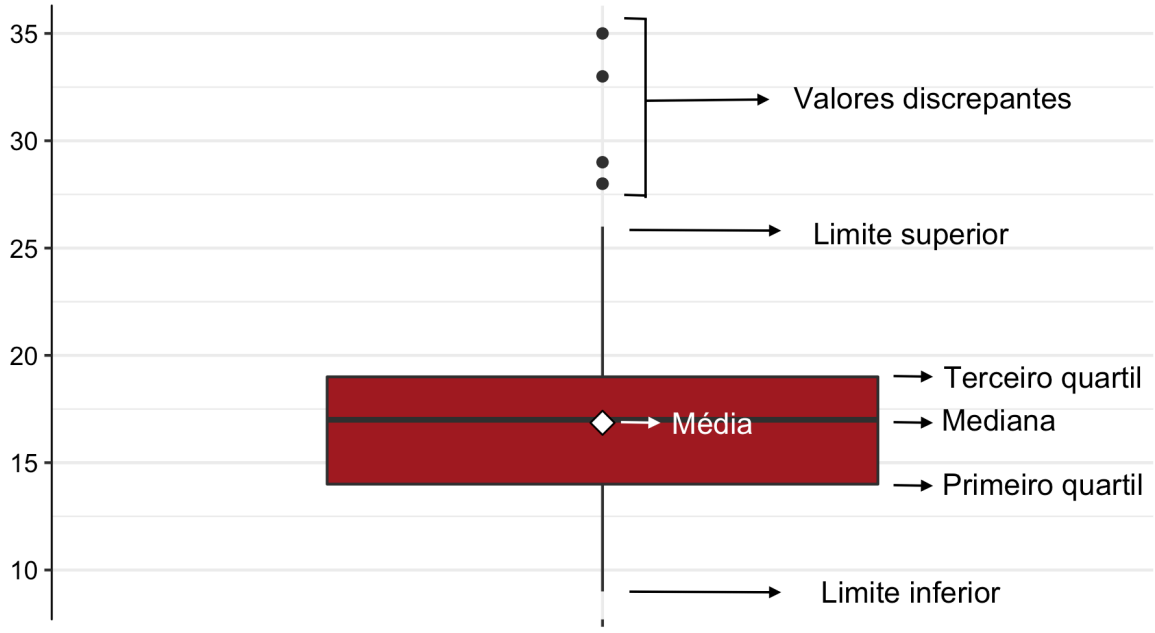

Figura 1: Exemplo de boxplot

A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

2.10 Histograma

O histograma é uma representação gráfica utilizada para a visualização da distribuição dos dados e pode ser construído por valores absolutos, frequência relativa ou densidade. A figura abaixo ilustra um exemplo de histograma.

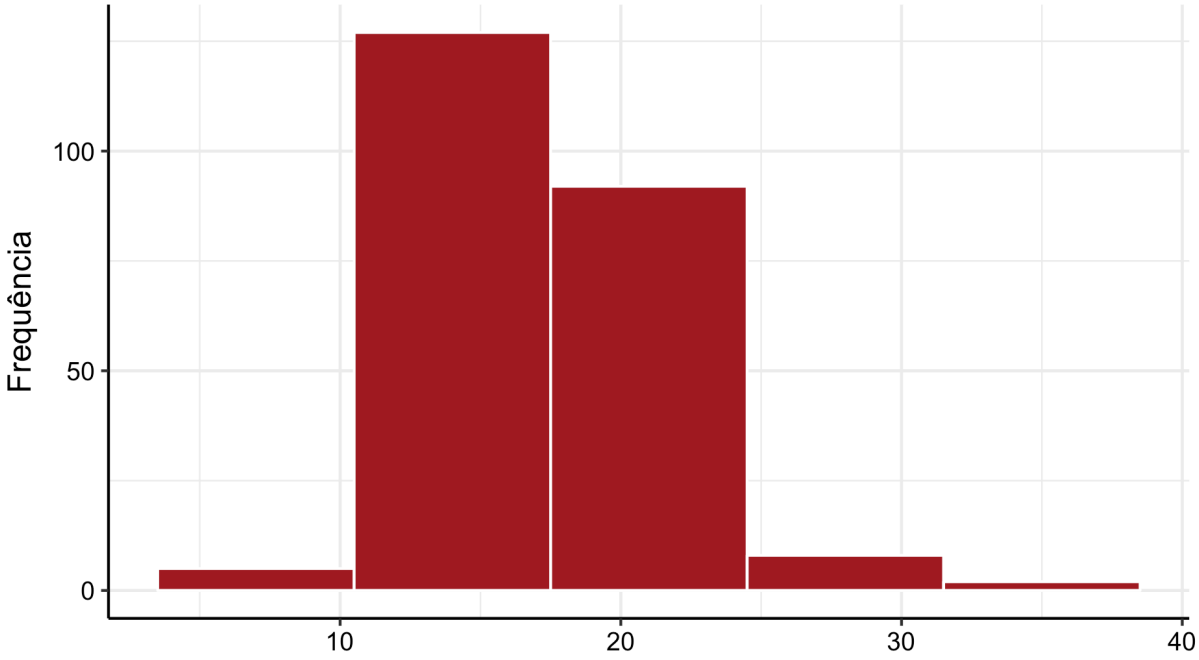

Figura 2: Exemplo de histograma

2.11 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

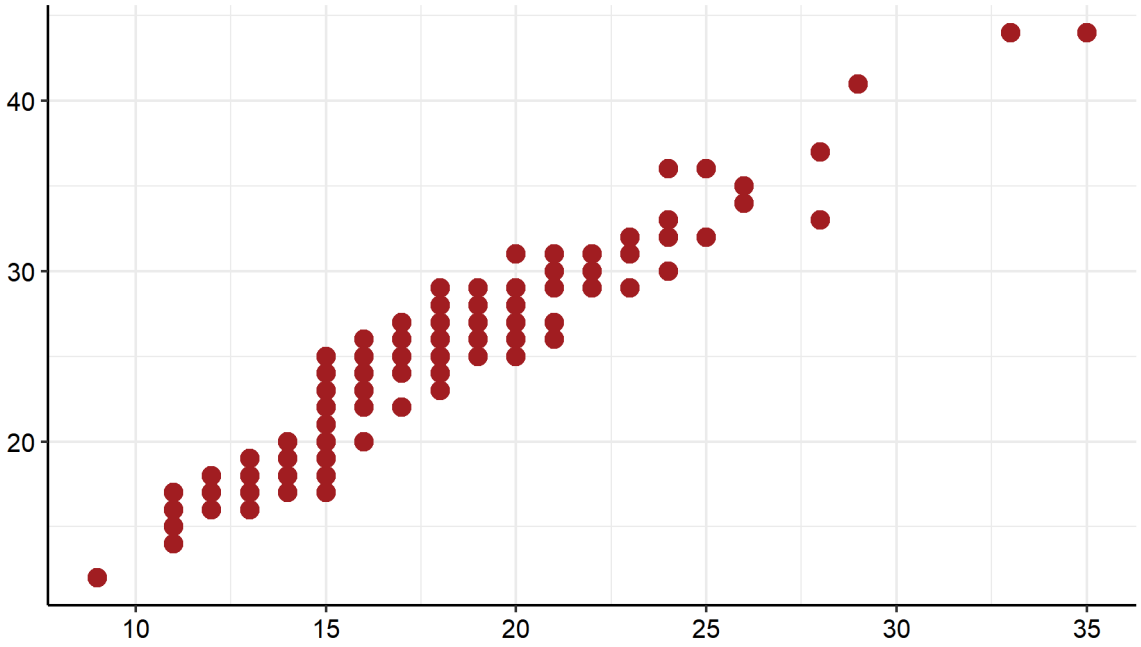

Figura 3: Exemplo de Gráfico de Dispersão

2.12 Tipos de Variáveis

2.12.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- **Nominais**: quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)
- **Ordinais**: quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

2.12.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- **Discretas**: quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- **Contínuas**: quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

3 Top 5 países com maior número de mulheres medalhistas

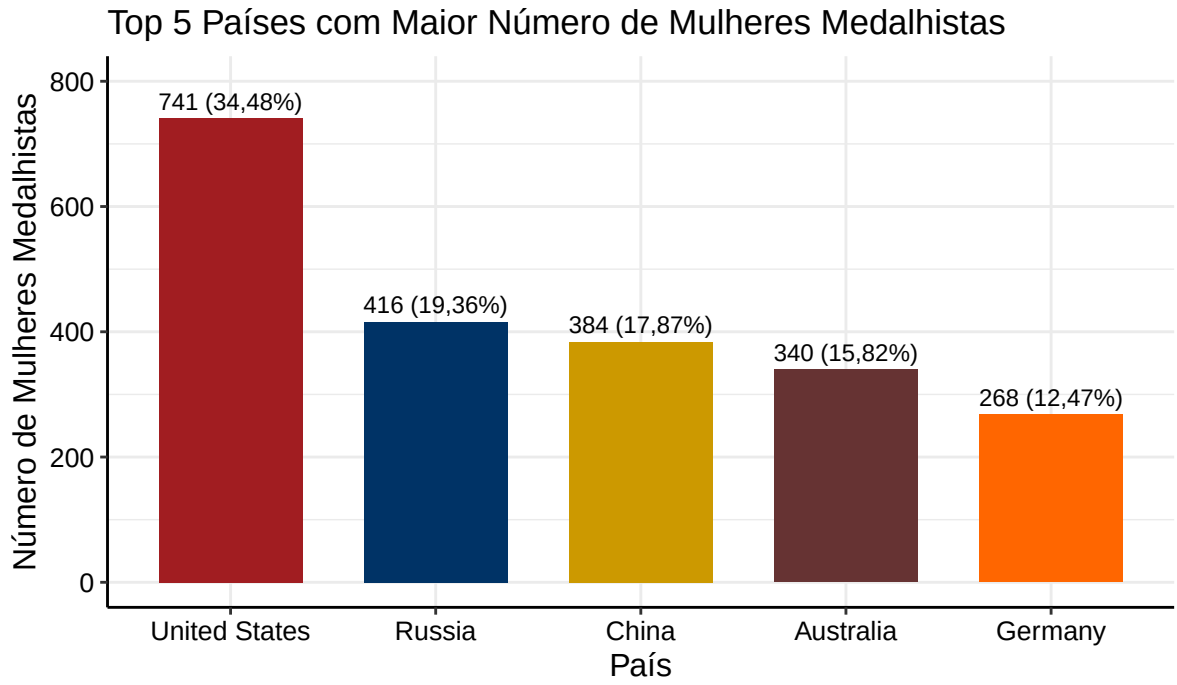

Figura 4: Gráfico de barras dos 5 países com maior número de mulheres medalhistas

Entre os anos de 2000 a 2016, os Estados Unidos se destacaram como o país com o maior número de mulheres medalhistas nas Olimpíadas, conquistando 741 medalhas, o que representa 34,48% do total. Em segundo lugar, a Rússia registrou 416 medalhas femininas, correspondendo a 19,36% do total. A China ocupa a terceira posição com 384 medalhas, o que equivale a 17,87% do total. A Austrália, com 340 medalhas (15,82%), é o quarto colocado. Por fim, a Alemanha completa o top 5 com 268 medalhas, representando 12,47% do total.

4 Valor IMC por esporte, estes sendo, ginástica, futebol, judô, atletismo e badminton

5 Top 3 medalhistas gerais por quantidade de cada tipo de medalha

6 Variação Peso por Altura

7 Conclusões