House of excellence

Consultores Responsáveis:
Francisco Italo Rio Andrade

Requerente:
João Vitor Neves

Brasília, 3 de novembro 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.

Para este relatório, será usada variáveis como país, sexo, idade, altura, peso, esporte, tipo de medalha e edição olímpica. A primeira análise identificará os cinco países com maior número de mulheres medalhistas, classificando-os em ordem decrescente de conquistas femininas. Já a segunda análise calculará o IMC de atletas em atletismo, badminton, futebol, ginástica e judô, para comparar a variação do índice entre esportes e verificar diferenças significativas, aplicando a ANOVA para essa comparação. Em seguida, a análise dos três maiores medalhistas em quantidade total de medalhas avaliará a frequência de medalhas de ouro, prata e bronze conquistadas por cada um e as associações entre o tipo de medalha e o atleta. Para entender a relação entre peso e altura, será realizada uma regressão linear, investigando se há correlação positiva, negativa ou inexistente entre essas variáveis. Essas análises combinam métodos descritivos, testes de hipóteses e regressão.

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 Frequência Relativa

A frequência relativa é utilizada para a comparação entre classes de uma variável categórica com c categorias, ou para comparar uma mesma categoria em diferentes estudos.

A frequência relativa da categoria j é dada por:

$$f_j = \frac{n_j}{n}$$

Com:

- $j = 1, ..., c$
- $n_j =$ número de observações da categoria j
- $n =$ número total de observações

Geralmente, a frequência relativa é utilizada em porcentagem, dada por:

$$100 \times f_j$$

2.2 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

2.3 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:

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

2.4 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.

2.5 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.

2.5.1 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

2.5.2 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

2.6 Desvio Padrão

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

2.6.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

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

Com:

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

2.6.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

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

Com:

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

2.7 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

2.8 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

2.9 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).

2.10 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.

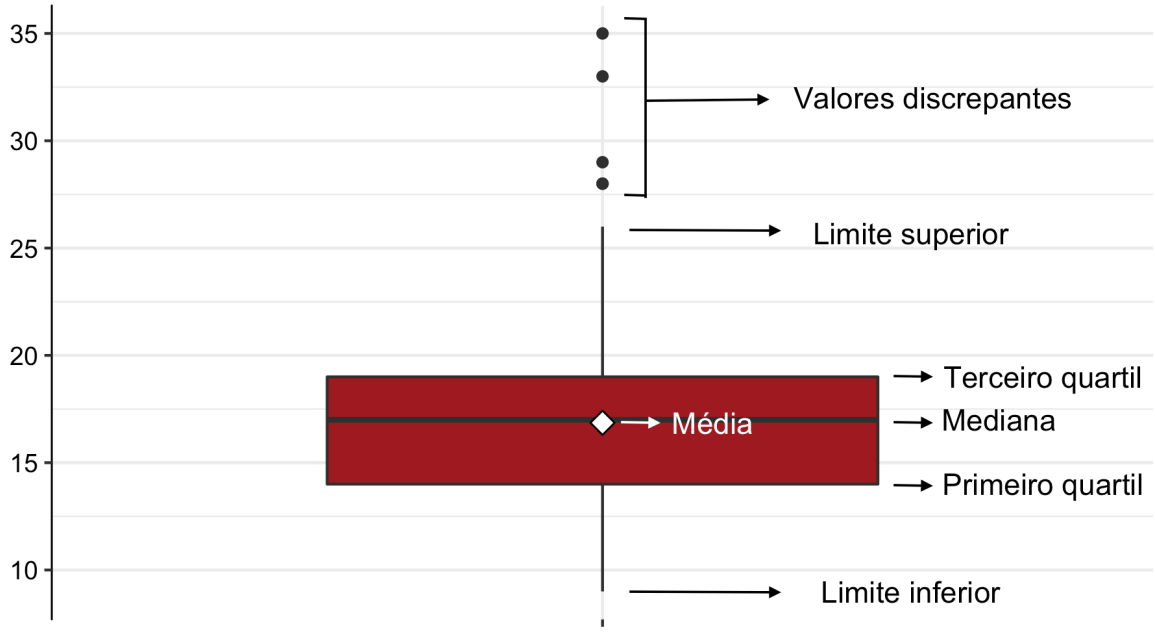


Figura 1: Exemplo de boxplot

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.

2.11 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.

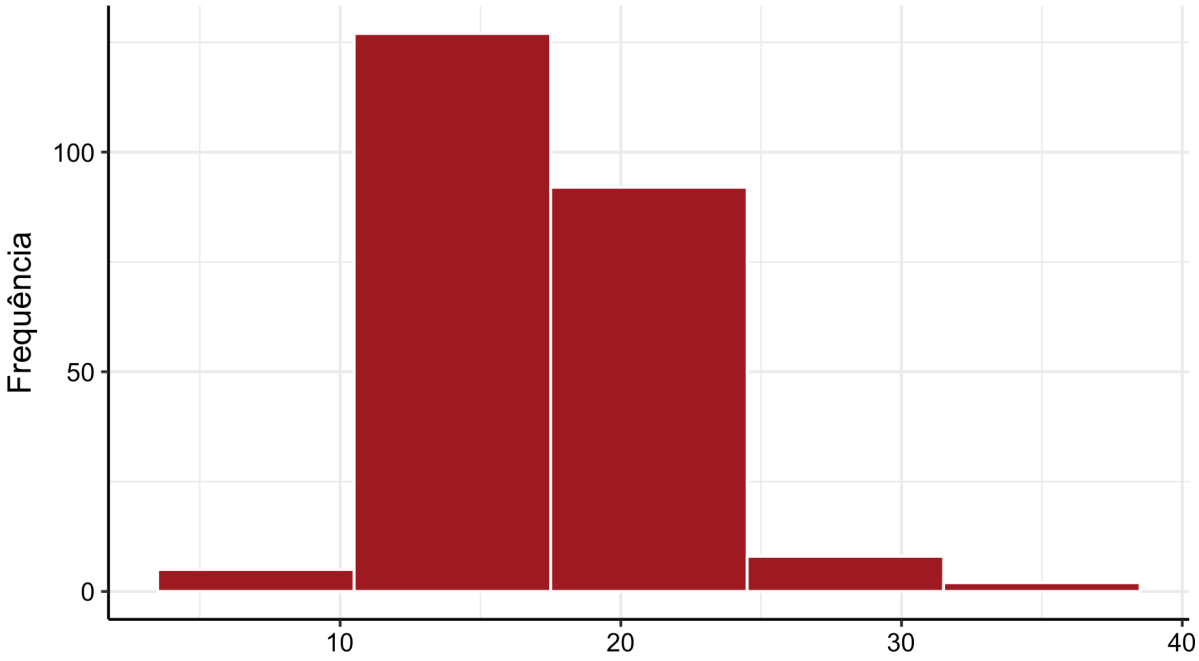


Figura 2: Exemplo de histograma

2.12 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.

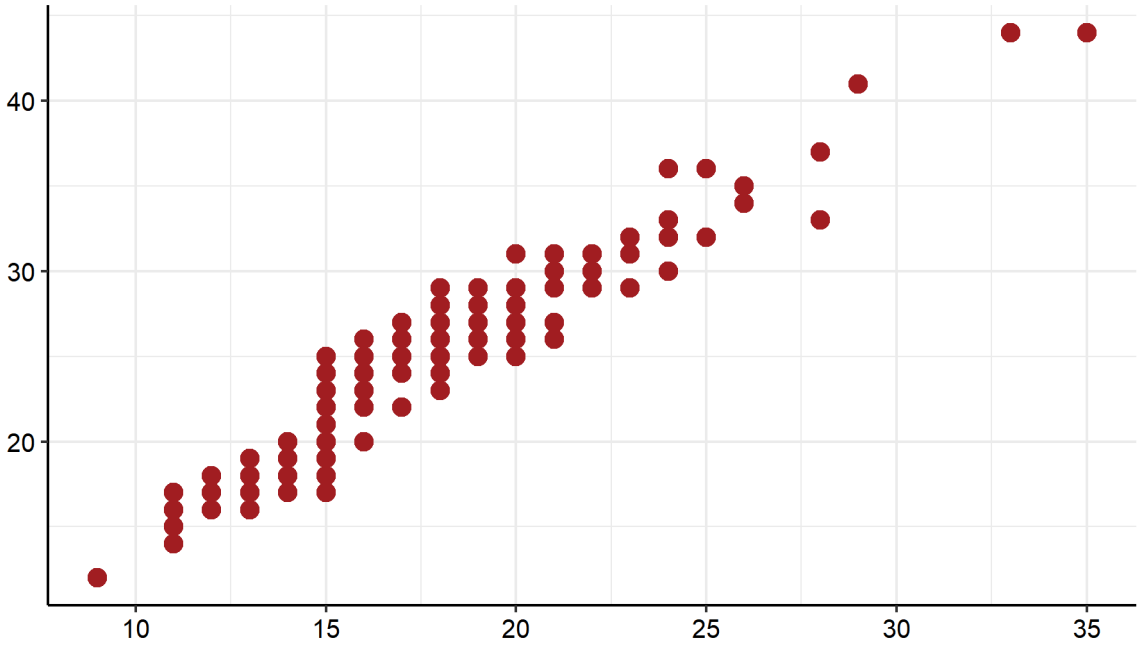


Figura 3: Exemplo de Gráfico de Dispersão

2.13 Tipos de Variáveis

2.13.1 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)

2.13.2 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)

2.14 Teste de Normalidade de Shapiro-Wilk

O **Teste de Shapiro-Wilk** é utilizado para verificar a aderência de uma variável quantitativa ao modelo da Distribuição Normal, sendo mais recomendado para amostras pequenas. A suposição de normalidade é importante para a determinação do teste a ser utilizado. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

A amostra deve ser ordenada de forma crescente para que seja possível obter as estatísticas de ordem. A estatística do teste é dada por:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)}\right)\right]$$

Com:

- K aproximadamente $\frac{n}{2}$
- $X_{(i)} =$ estatística de ordem *i*
- $D = \sum_{i=1}^{n}(X_i - \bar{X})^2$, em que $\bar{X}$ é a média amostral
- $a_i = $ constantes que apresentam valores tabelados

2.15 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é utilizado para comparar dois ou mais grupos independentes sem supor nenhuma distribuição. É um método baseado na comparação de postos, os quais são atribuídos a cada observação de uma variável quantitativa após serem ordenadas.

As hipóteses do teste de Kruskal-Wallis são formuladas da seguinte maneira:

$$\begin{cases} H_0 : \text{Não existe diferença entre os grupos} \\ H_1 : \text{Pelo menos um grupo difere dos demais} \end{cases}$$

A estatística do teste de Krukal-Waliis é definida da seguinte maneira:

$$H_{Kruskal-Wallis} = \frac{\left[\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n+1)}{1 - \left[\frac{\sum_j (t_j^3 - t_j)}{n^3 - n}\right]} \approx \chi^2_{(k-1)}$$

Com: - $k =$ número de grupos

- $R_i =$ soma dos postos do grupo i
- $n_i =$ número de elementos do grupo i
- $n =$ tamanho total da amostra
- $t_j =$ número de elementos no j-ésimo empate (se houver)

Se o p-valor for menor que o nível de significância α, rejeita-se a hipótese nula.

2.16 Teste de Normalidade de Anderson-Darling

O **teste de Normalidade de Anderson-Darling** é utilizado para verificar se uma amostra aleatória $X_1, X_2, ..., X_n$ de uma variável quantitativa segue uma distribuição Normal de probabilidade ou não. O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula for verdadeira, espera-se que o p-valor esteja acima do nível de significância α.

3 Análises

3.1 Top 5 países com maior número de mulheres medalhistas

A análise, buscamos identificar os países que tiveram o maior número de mulheres medalhistas em Jogos Olímpicos de 2000 a 2016. Utilizou-se dados das edições de Sydney 2000, Atenas 2004, Pequim 2008, Londres 2012 e Rio 2016, considerando as variáveis de sexo (feminino), país de origem das atletas (Team), e a presença de uma medalha (Medal).A variavel sexo é uma variavel qualitativa nominal, já a variável O objetivo principal foi compreender quais nações se destacaram em termos de conquistas femininas, utilizando como métricas o número total de medalhistas por país e a frequência relativa de cada país em relação ao total de mulheres medalhistas.

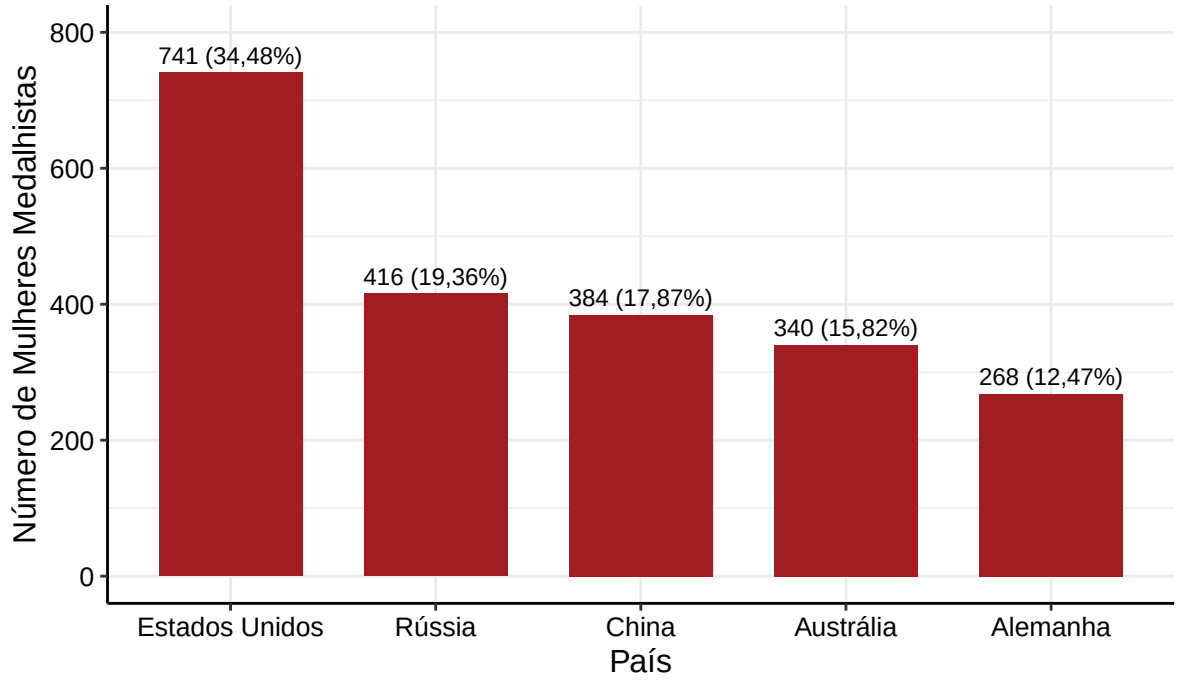


Figura 4: Gráfico de barras dos 5 países com maior número de mulheres medalhistas

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.

3.2 Análise do IMC para os esportes selecionados

Nesta análise, busca-se comparar o Índice de Massa Corporal (IMC) de atletas olímpicos que competiram em diferentes modalidades esportivas, especificamente Atletismo, Badminton, Futebol, Ginástica e Judô. O objetivo é identificar se há diferenças significativas nos valores de IMC entre esses esportes, bem como entender quais esportes tendem a ter IMCs mais altos ou mais baixos.

$$IMC = \frac{Peso(kg)}{Altura(m^2)}$$

Em termos gerais, os valores de IMC podem ser categorizados da seguinte forma:

- Abaixo de 18,5: Peso abaixo do ideal.
- 18,5 a 24,9: Peso normal ou saudável.
- 25,0 a 29,9: Sobrepeso.
- 30,0 a 34,9: Obesidade grau I.
- 35,0 a 39,9: Obesidade grau II.
- Acima de 40,0: Obesidade grau III.

Tabela 1: Tabela das estatísticas descritivas por esporte

Esporte	Tamanho Amostral	Média	Mediana	Desvio padrão	Min	Max
Atletismo	939	22,30296	21,44755	3,862608	15,82215	44,37866
Badminton	120	22,21188	22,28257	1,503154	18,93699	26,72991
Futebol	513	22,50681	22,49133	1,729404	16,72767	29,06875
Ginástica	348	20,68349	21,09373	2,380924	15,15965	26,44626
Judô	280	25,69914	24,67548	5,121888	18,51779	56,49531

No Atletismo, a média do IMC é de 22,3, com uma mediana de 21,4 e um desvio padrão de 3,86. Isso sugere que os valores de IMC dos atletas são relativamente próximos, mas há alguns casos extremos, como evidenciado pelo valor máximo de 44,4.

Os atletas de Badminton apresentam uma média de IMC de 22,2, e uma mediana de 22,3, com um desvio padrão baixo de 1,50. Indicando uma menor variabilidade nos valores de IMC, com a maioria dos atletas concentrados em torno da média.

Os jogadores de Futebol apresentam uma média de IMC ligeiramente maior, de 22,5, e uma mediana de 22,5, com um desvio padrão baixo de 1,73. Isso indica uma menor variabilidade nos valores de IMC, com a maioria dos jogadores concentrados em torno da média.

Na Ginástica, o IMC médio é de 20,7, enquanto a mediana é de 21,1, e o desvio padrão é de 2,38, indicando também uma baixa variabilidade e uma distribuição mais homogênea dos valores de IMC entre os atletas.

Já os atletas de Judô apresentam a maior média de IMC, de 25,7, com uma mediana de 24,7 e um desvio padrão mais elevado, de 5,12. Esse maior desvio padrão sugere uma variabilidade significativa no IMC dos judocas, com um valor máximo de 56.5, indicando que alguns atletas possuem um IMC substancialmente maior, o que pode estar relacionado à necessidade de uma maior massa corporal nesse esporte de combate.

De acordo com os dados tem se os judocas tendem a ter um IMC mais elevados, enquanto os ginastas apresentam os menores. As diferenças na variabilidade entre os grupos refletem os distintos requisitos físicos de cada esporte, como força e agilidade.

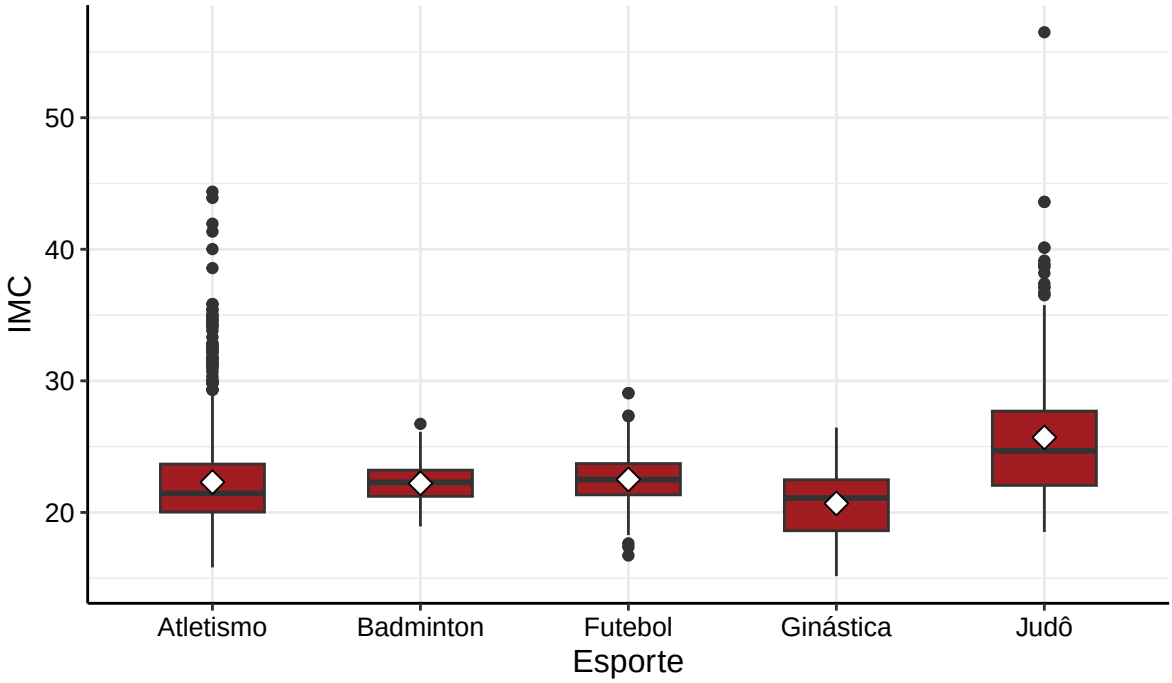


Figura 5: Boxplot da Comparação do IMC entre Esportes

Observa-se que os atletas de Atletismo e Badminton tendem a ter uma mediana de IMC menor em comparação aos de Ginástica e Judô, sugerindo que essas duas modalidades têm atletas com IMC menores. O Judô apresenta a maior mediana entre os esportes analisados, indicando uma tendência a ter um IMC mais elevados entre seus atletas.

O Judô se destaca por ter uma caixa maior, sugerindo mais variação no IMC dos

seus atletas em comparação aos outros esportes. Por outro lado, os da Ginástica é menor, o que indica uma menor variação no IMC dos atletas dessa modalidade.

O boxplot também exibe varios pontos fora dos limites, também conhecidos com outliers, que representam valores atípicos. Esportes como Atletismo e Judô apresentam uma quantidade significativa de outliers, com alguns atletas tendo IMC consideravelmente mais altos que a maioria de seus pares na mesma modalidade.

3.2.1 Normalidade

O teste de Shapiro-Wilk foi aplicado para cada esporte para verificar se os dados de IMC seguem uma distribuição normal.

Tabela 2: Teste de Shapiro-Wilk

P-valor
8,537905e-28
6,391550e-01
1,895904e-03
9,538378e-05
7,093824e-15

$\begin{cases} H_0 : \text{A distribuição dos dados de IMC para cada esporte é normal.} \\ H_1 : \text{A distribuição dos dados de IMC para cada esporte não é normal.} \end{cases}$ \

Os valores-p resultantes para todos os esportes foram extremamente baixos, menores que um nível de significância preestabelecido de 5%, indicando que os dados não seguem uma distribuição normal.

3.2.2 Teste de Kruskal-Wallis

Dada a violação da normalidade, foi aplicado o teste de Kruskal-Wallis, que é um teste não-paramétrico.

Tabela 3: Teste de Kruskal-Wallis

	Estatística	P-valor	Grupos	Tamanho amostral
Kruskal-Wallis chi-squared	1984,408	0e+00	36	67474

$\begin{cases} H_0 : \text{Não há diferença significativa nas distribuições de IMC entre as modalidades esportiv} \\ H_1 : \text{Há pelo menos uma diferença significativa nas distribuições de IMC entre as modalid} \end{cases}$

O resultado do p-valor foi menor que 5%, indicando diferenças significativas do IMC entre os grupos de esportes.

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

A análise tem como objetivo identificar os três atletas com mais medalhas nas Olimpíadas de 2000 (Sydney), 2004 (Atenas), 2008 (Pequim), 2012 (Londres) e 2016 (Rio), além de examinar a quantidade de cada tipo de medalha (ouro, prata e bronze) conquistada por eles. Também será investigada, por meio de métodos estatísticos como o teste qui-quadrado de independência, a existência de uma relação entre o atleta e o tipo de medalha.

Tabela 4: Tabela dos Medalhistas e Total de Medalhas

Nome	Total de Medalhas
Michael Fred Phelps, II	28
Natalie Anne Coughlin (-Hall)	12
Ryan Steven Lochte	12

Com base na tabela, identifica-se que os três atletas com o maior número de medalhas em todas as edições analisadas foram: Michael Fred Phelps, II, com um total de 28 medalhas; Natalie Anne Coughlin, com 12 medalhas; e Ryan Steven Lochte, também com 12 medalhas.

Tabela 5: Tabela da Distribuição das Medalhas

Nome	Medalha	Quantidade
Michael Fred Phelps, II	Bronze	2
Michael Fred Phelps, II	Ouro	23
Michael Fred Phelps, II	Prata	3
Natalie Anne Coughlin (-Hall)	Bronze	5
Natalie Anne Coughlin (-Hall)	Ouro	3
Natalie Anne Coughlin (-Hall)	Prata	4
Ryan Steven Lochte	Bronze	3
Ryan Steven Lochte	Ouro	6
Ryan Steven Lochte	Prata	3

A distribuição das medalhas revelou-se diferenças significativas entre os atletas. Michael Phelps, que se destacou como o maior medalhista, conquistou 23 medalhas

de ouro, 3 de prata e 2 de bronze, tendo um desempenho excepcional em conquistas medalhas de ouro. Natalie Coughlin, por outro lado, teve uma distribuição mais equilibrada, com 3 medalhas de ouro, 4 de prata e 5 de bronze. Já Ryan Lochte obteve 6 medalhas de ouro, 3 de prata e 3 de bronze, também apresentando um perfil balanceado em termos de tipo de medalha.

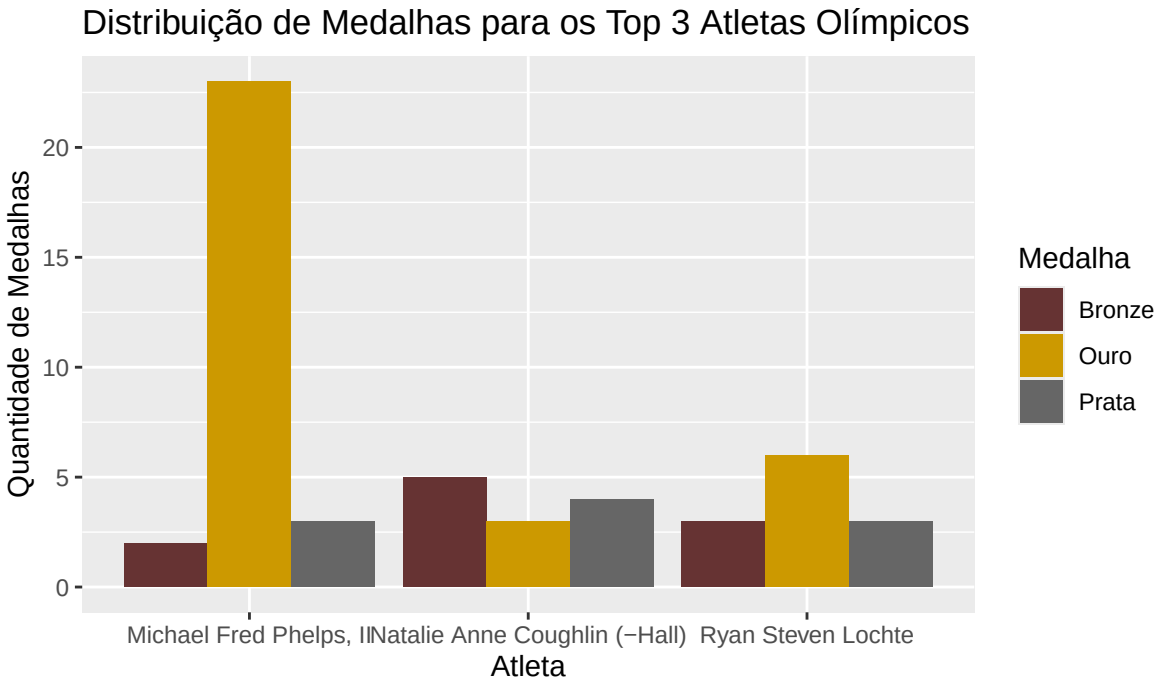


Figura 6: Gráfico de barras da distribuição das medalhas

A distribuição das medalhas foi visualizada em um gráfico de barras que evidenciou o predomínio de medalhas de ouro nas conquistas de Michael Phelps, enquanto Natalie Coughlin e Ryan Lochte apresentaram uma proporção mais distribuída entre os diferentes tipos de medalha.

Podemos confirmar matematicamente fazendo um teste de independência.

$$\begin{cases} H_0 : \text{Não há relação entre o medalhista e o tipo de medalha conquistada. Em outras palav} \\ H_1 : \text{Existe uma relação entre o medalhista e o tipo de medalha conquistada. Isso significa} \end{cases}$$

Tabela 6: Resultados do teste qui-quadrado

	Estatítica	Graus de Liberdade	P-valor
X-squared	12,7756	4	0,012426

Para verificar se existia uma relação estatisticamente significativa entre os atletas e os tipos de medalhas, realizo-se um teste qui-quadrado de independência. Os resultados do teste indicaram um valor de X-quadrado de 12,776 com 4 graus de liberdade

e um valor-p de 0,01243. Este valor-p, sendo inferior ao nível de significância de 5%, que levou a rejeitar a hipótese nula de independência entre as variáveis. Ou seja, os resultados indicam que existe uma relação significativa entre os atletas analisados e os tipos de medalhas que conquistaram. Isso sugere que a distribuição das medalhas de ouro, prata e bronze não ocorre de forma uniforme entre os três atletas, havendo diferenças marcantes.

3.4 Variação Peso por Altura

A análise realizada tem como objetivo compreender a relação entre o peso e a altura dos atletas, investigando se existe uma correlação direta entre essas variáveis. Em outras palavras, a análise busca responder se, à medida que o peso aumenta, a altura dos atletas também tende a aumentar, ou se não há uma relação clara entre essas duas características físicas. Para essa análise, são utilizadas as variáveis peso e altura, ambas sendo variáveis quantitativas contínuas.

Tabela 7: Tabela das estatísticas descritivas da Altura dos Atletas

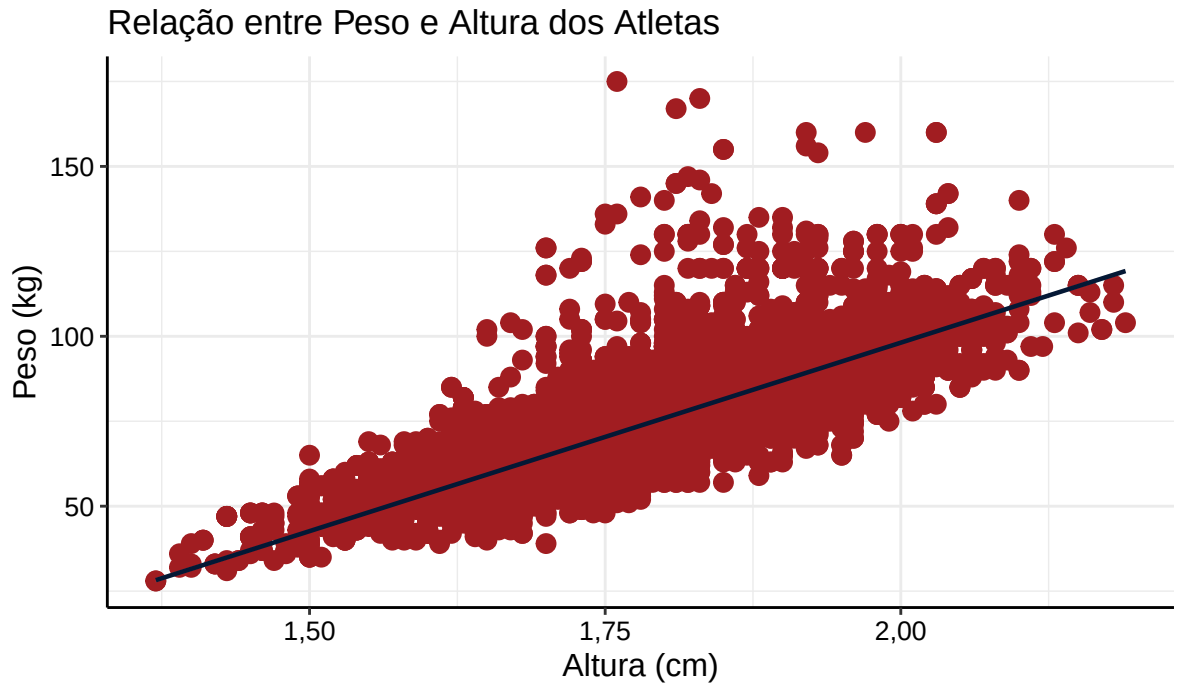
Altura	Tamanho Amostral	Média	Mediana	Desvio padrão	Min	Max
[1,3,1,4)	7	16,87035	16,56228	1,838715	14,91820	19,89794
[1,4,1,5)	52	19,61145	19,50058	2,679345	15,15965	23,87278
[1,5,1,6)	385	21,17484	21,05169	2,484375	15,35019	28,88887
[1,6,1,7)	2308	21,85983	21,56453	2,755816	13,49480	43,59858
[1,7,1,8)	2623	22,65468	22,38630	3,033104	15,67346	56,49531
[1,8,1,9)	3007	23,81068	23,29121	3,267657	16,65448	50,97520
[1,9,2)	1183	24,52185	24,45757	2,739306	17,09400	43,40274
[2,2,1)	358	24,78323	24,74312	2,729724	19,30644	38,82644
[2,1,2,2)	34	24,80092	24,87829	2,610155	20,40815	31,74601

A Tabela 7 observa-se que a média e a mediana da altura aumentam conforme os intervalos se expandem, sugerindo uma tendência de crescimento no porte físico dos atletas nos intervalos superiores. O desvio padrão também é maior nos grupos com atletas mais altos, indicando que há uma maior variação de altura entre esses competidores. Além disso, a diferença entre os valores mínimos e máximos em cada faixa evidencia a diversidade de estaturas, mostrando que atletas de alturas variadas competem em diferentes esportes, dependendo das exigências físicas de cada modalidade.

Tabela 8: Tabela das estatísticas descritivas do Peso dos Atletas

Peso	Tamanho Amostral	Média	Mediana	Desvio padrão	Min	Max
[20,30)	2	14,91820	14,91820	0,000000	14,91820	14,91820
[30,40)	53	16,86067	16,89188	1,307309	13,49480	20,11970
[40,50)	404	18,73121	18,51779	1,569327	14,88094	22,98399
[50,60)	1686	20,38348	20,41521	1,465597	16,27883	25,77776
[60,70)	2473	22,01284	21,95246	1,545138	17,09400	28,88887
[70,80)	2333	23,18689	23,08252	1,656988	18,74218	29,73388
[80,90)	1448	24,62049	24,45355	1,791124	20,22604	32,38833
[90,100)	1027	25,93563	25,50758	2,025211	21,29070	36,73092
[100,110)	330	27,64852	26,79789	2,903483	21,66109	37,46553
[110,120)	116	30,37240	29,82054	3,542865	24,19828	40,83042
NA	85	37,24559	36,56507	5,706345	26,89059	56,49531

A Tabela 8 fornece uma análise focada nos intervalos de peso. Assim como na altura, a média e a mediana do peso dos atletas aumentam de forma progressiva ao longo dos intervalos, refletindo a categorização por faixas de peso. Os valores de desvio padrão sugerem uma maior variação nos grupos intermediários, onde há uma maior diversidade de composição corporal. A diferença entre os valores mínimos e máximos dentro de cada intervalo também mostra a diversidade de pesos, o que pode ser um reflexo das exigências físicas de diferentes esportes, onde atletas podem precisar de mais ou menos massa corporal para obter desempenho ideal.



Pode-se observar uma correlação positiva entre altura e peso, onde, à medida que a altura dos atletas aumenta, o peso tende a aumentar também. Sugerindo que atletas mais altos geralmente têm um peso. A dispersão dos pontos em torno da linha de tendência mostra, no entanto, uma variação considerável. Atletas de mesma altura apresentam pesos diferentes, o que pode refletir a diversidade de esportes e de composição corporal, onde fatores como a quantidade de massa muscular e gordura influenciam significativamente o peso.

Tabela 9: Tabela da normalidade da Altura dos Atletas

	Estatística	Valor P
A	8,801266	0

Tabela 10: Tabela da normalidade do Peso dos Atletas

	Estatística	Valor P
A	52,7453	0

$$\begin{cases} H_0 : \text{Os dados vêm de uma distribuição normal} \\ H_1 : \text{Os dados não vêm de uma distribuição normal} \end{cases} \backslash$$

Para verificar se essas variáveis seguem uma distribuição normal, foi realizado o Teste de Anderson-Darling, que é utilizado para avaliar a normalidade dos dados de grandes amostras. Tanto para a altura quanto para o peso, os resultados apontaram p-valores extremamente baixos, o que leva à rejeição da hipótese nula de normalidade. Isso indica que tanto o peso quanto a altura dos atletas não seguem uma distribuição normal, sendo distribuídos de maneira assimétrica.

Tabela 11: Teste de Bartlett para homogeneidade de variâncias entre grupos de peso

	Estatística	P-valor
Bartlett's K-squared	Inf	0

$$\begin{cases} H_0 : \text{As variâncias dos grupos são iguais} \\ H_1 : \text{Pelo menos uma das variâncias é diferente} \end{cases} \backslash$$

Tabela 12: Teste de Levene para homogeneidade de variâncias entre grupos de peso

	Df	F value	Pr(>F)
group	9	22,05735	0
	9862	NA	NA

$$\begin{cases} H_0 : \text{As variâncias dos grupos são iguais} \\ H_1 : \text{Pelo menos uma das variâncias é diferente} \end{cases} \backslash$$

O Teste de Bartlett e o Teste de Levene foram conduzidos e ambos indicaram resultados significativos, com os p-valores extremamente baixos. Isso sugere que as variâncias entre os grupos de diferentes faixas de peso não são homogêneas, ou seja, há uma diferença significativa nas variâncias entre os grupos, o que pode afetar a análise de correlação entre as variáveis.

Tabela 13: Resultados da Regressão Linear Simples para Altura em função do Peso

	Estimativa	Erro Padrão	Estatística t	P-valor
(Intercept)	1,3499555	0,0032677	413,1258	0
Peso	0,0058458	0,0000431	135,5422	0

Para investigar a relação entre peso e altura, foi realizada uma análise de regressão linear, onde a altura foi ajustada em função do peso. O modelo ajustado revelou uma equação onde a altura aumenta conforme o peso aumenta. O coeficiente de regressão associado ao peso foi positivo e estatisticamente significativo, reforçando a ideia de que há uma relação positiva entre as duas variáveis.

4 Conclusões

A primeira análise dos países com o maior número de mulheres medalhistas nos Jogos Olímpicos de 2000 a 2016 revelou um destaque claro dos Estados Unidos, que lideram com uma grande margem, possuindo 741 medalhistas, o que representa 34,48% do total. Em seguida, a Rússia ocupa a segunda posição com 416 medalhistas (19,36%), a China aparece logo atrás, com 384 medalhistas (17,87%). Austrália e Alemanha completam o Top 5, com 340 (15,82%) e 268 (12,47%) medalhistas, respectivamente.A diferença expressiva entre os Estados Unidos e os demais países sugere uma forte tradição e investimento no esporte feminino, refletindo-se em um número maior de atletas de alto nível. Por outro lado, Rússia e China também se destaquem, a diferença na quantidade de medalhistas são significativamente menores quando comparadas aos Estados Unidos.

Na segunda análise observou-se diferenças significativas no IMC entre as modalidades esportivas analisadas (Atletismo, Badminton, Futebol, Ginástica e Judô). Devido à não normalidade dos dados, foi aplicado o teste de Kruskal-Wallis, que confirmou essas diferenças. O Judô apresentou os maiores valores de IMC, refletindo um perfil corporal robusto, enquanto Ginástica e Atletismo mostraram IMCs mais baixos, associados à agilidade. Futebol e Badminton ficaram em posição intermediária, com valores de IMC superiores aos de Ginástica e Atletismo, mas abaixo dos do Judô.

Na terceira análise mostrou que há diferenças significativas na forma como os três principais medalhistas olímpicos conquistaram suas medalhas, com Phelps se destacando pelo número expressivo de ouros, enquanto Coughlin e Lochte apresentaram maior equilíbrio entre os diferentes tipos de medalha. Essas diferenças são estatisticamente relevantes, conforme evidenciado pelo teste qui-quadrado, e refletem as particularidades de desempenho de cada atleta ao longo de suas participações olímpicas. Esse entendimento pode contribuir para análises futuras sobre o impacto de diferentes fatores (como modalidade esportiva, preparo físico e características pessoais) no desempenho dos atletas olímpicos mais bem-sucedidos.

A quarta análise mostra que há uma correlação positiva significativa entre o peso e a altura dos atletas.