Análise do Desempenho e Perfil dos Atletas Olímpicos: Um Panorama das Conquistas e Características Físicas de 2000 a 2016

Consultores Responsáveis:
Francisco Ítalo Rios Andrade

Requerente:
João Vitor Neves

Brasília, 7 de novembro de 2024.

ESTAT
consultoria estatística

Sumário

Página

1 Introdução 3
2 Referencial Teórico 4
2.1 Frequência Relativa 4
2.2 Média 4
2.3 Mediana 5
2.4 Quartis 5
2.5 Variância 5
2.5.1 Variância Amostral 6
2.6 Desvio Padrão 6
2.6.1 Desvio Padrão Amostral 6
2.7 Coeficiente de Variação 6
2.8 Coeficiente de Assimetria 7
2.9 Coeficiente de Correlação de Pearson 7
2.10 Boxplot 8
2.11 Histograma 9
2.12 Gráfico de Dispersão 9
2.13 Tipos de Variáveis 10
2.13.1 Qualitativas 10
2.13.2 Quantitativas 10
2.14 Teste de Normalidade de Shapiro-Wilk 11
2.15 Teste de Kruskal-Wallis 11
2.16 Teste de Normalidade de Anderson-Darling 12
3 Análises 13
3.1 Top 5 países com maior número de mulheres medalhistas únicas 13
3.2 Análise do IMC para os esportes selecionados 14
3.2.1 Normalidade 16
3.2.2 Teste de Kruskal-Wallis 17
3.3 Top 3 medalhistas gerais por quantidade de cada tipo de medalha 17
3.4 Variação Peso por Altura 20
4 Conclusões 26

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.

O banco de dados foi fornecidos pelo cliente separado por paises sede da olimpiada e ano de sua realização e contém as seguintes variáveis: nome do atleta, sexo, idade, altura, peso, país que o atleta compete, esporte, evento e tipo de medalha. A primeira análise identifica os cinco países com maior número de mulheres medalhistas, classificando-os em ordem decrescente de conquistas femininas. Já a segunda análise calculou 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

Este relatório é composto por técnicas estatísticas que serão descritas a seguir de acordo com o que foi utilizado em tal estudo.

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 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 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 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x}) \, (y_i - \bar{y}) \right]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Onde:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $\bar{x} =$ média dos valores da variável X
- $\bar{y} =$ média dos valores da variável Y

Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

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.

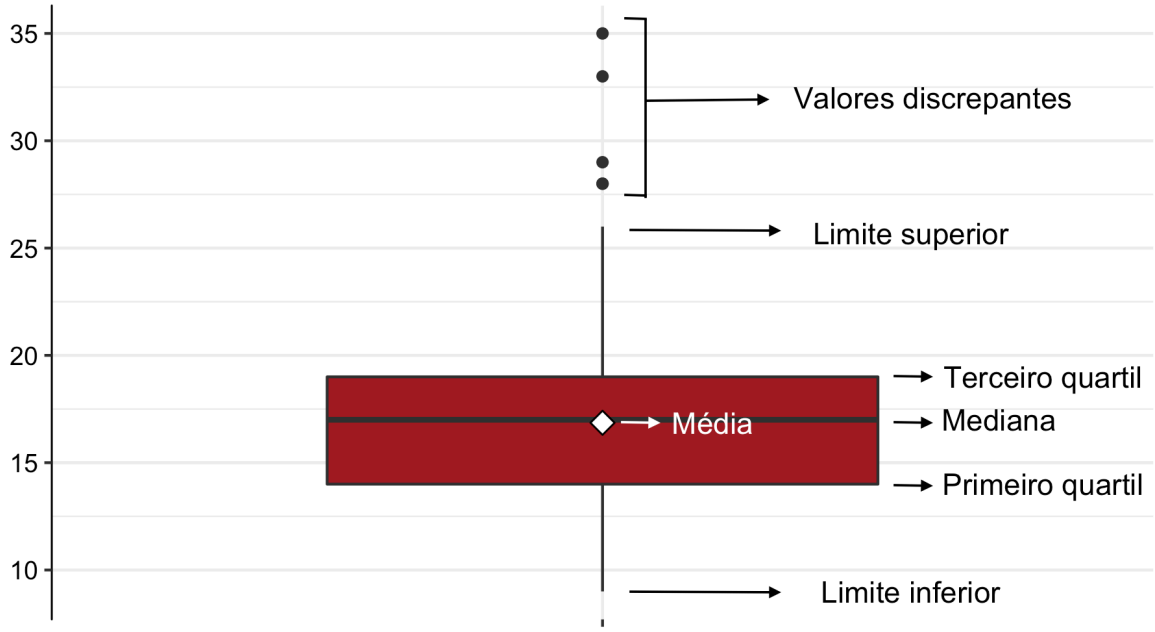


Figura 1: Exemplo de boxplot

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.

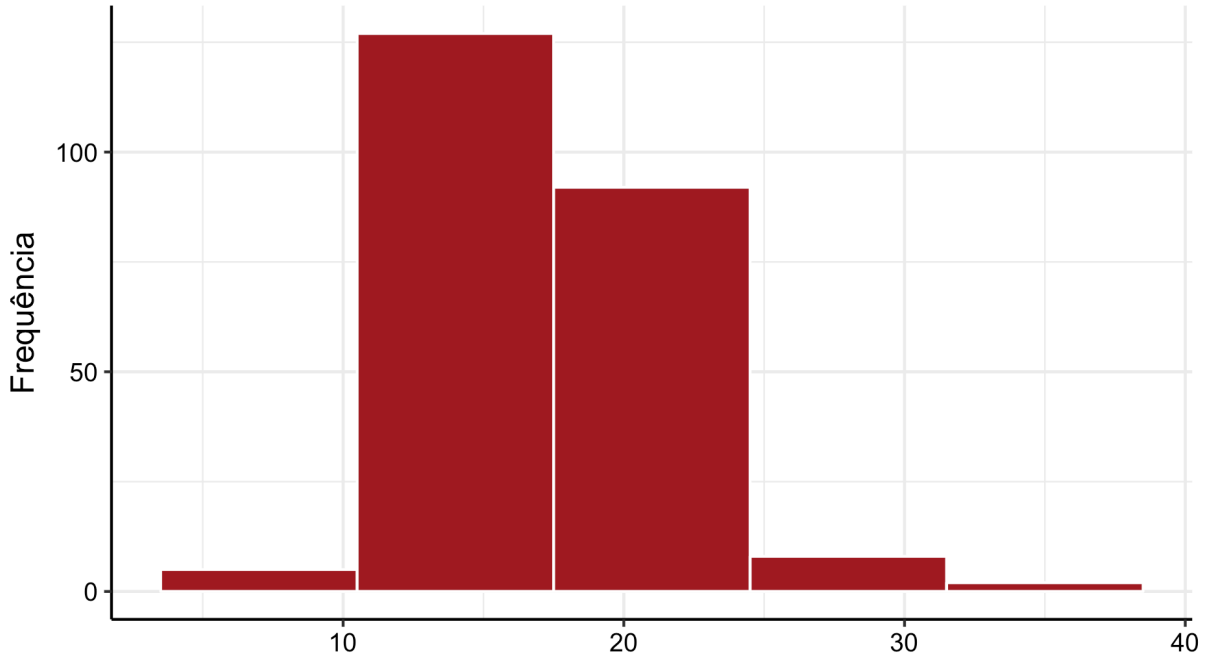


Figura 2: Exemplo de histograma

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.

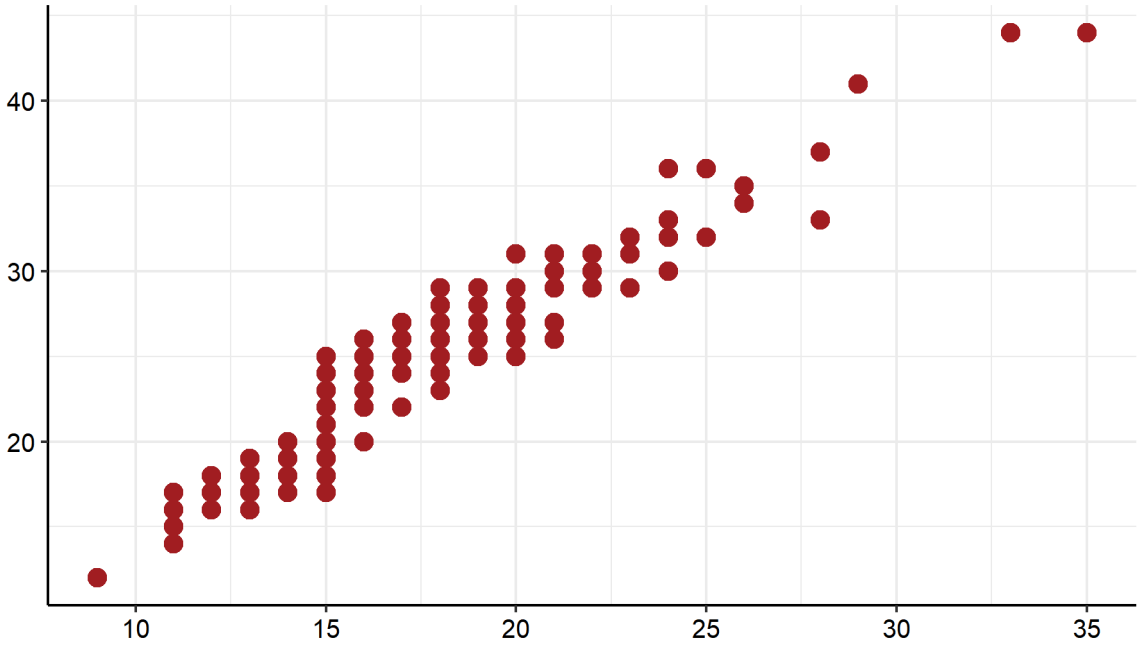


Figura 3: Exemplo de Gráfico de Dispersão

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 únicas

Na análise, busca-se identificar os países que tiveram o maior número de mulheres medalhistas nos 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 (Time), e a presença de uma medalha (Medalha). As três variáveis estudadas são variáveis qualitativas nominais. O objetivo principal foi compreender quais nações se destacaram em termos de conquistas femininas, sem duplicações de resultados, ou seja, apenas o número de mulheres medalhistas únicas. Utilizando como métricas o número total de medalhistas por país e a frequência relativa de cada país em relação ao total de mulheres medalhistas.

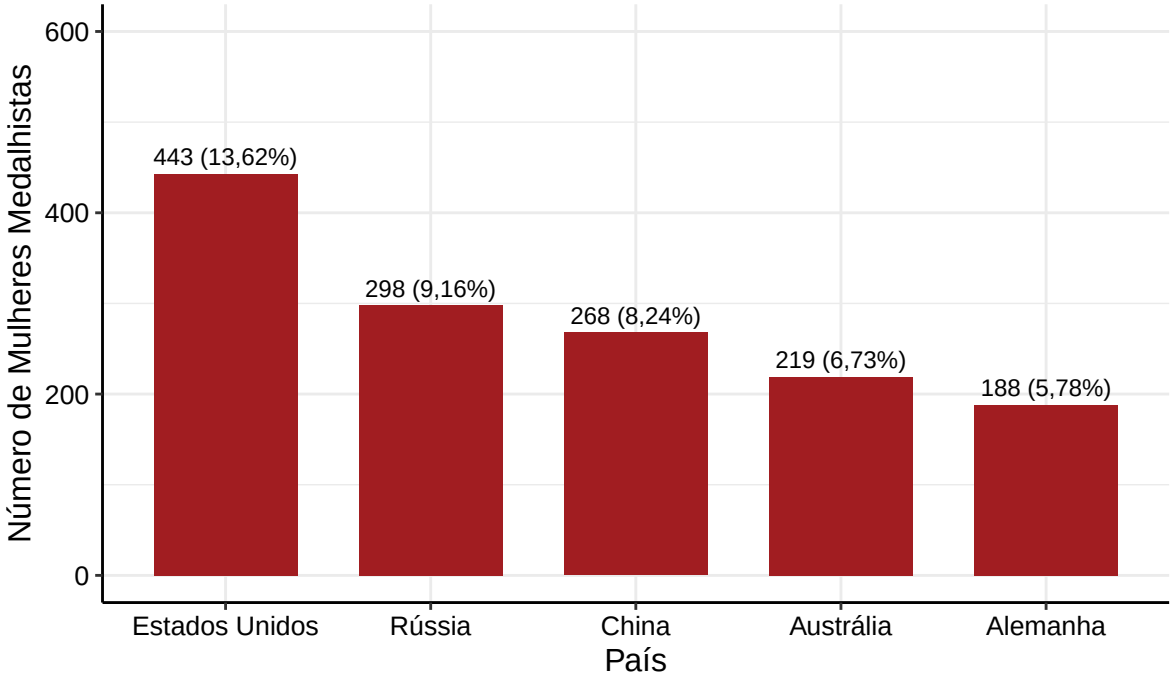


Figura 4: Gráfico de barras dos 5 países com maior número de mulheres medalhistas

Entre os anos de 2000 a 2016, os Estados Unidos se destacaram como o país com o maior número de mulheres medalhistas nas Olimpíadas, totalizando 443 atletas únicas, o que representa 13,62% do total de mulheres medalhistas analisadas. Em segundo lugar, a Rússia registrou 298 mulheres medalhistas, correspondendo a 9,16% do total. A China ocupa a terceira posição com 268 medalhistas, o que equivale a 8,24% do total. A Austrália, com 219 atletas únicas (6,73%), é o quarto colocado. Por fim, a Alemanha completa o top 5 com 188 medalhistas, representando 5,78% 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. As variáveis trabalhadas foram as seguintes: esporte é uma variável qualitativa nominal, altura variável quantitativa contínua e peso é uma variável quantitativa contínua. Depois de calculado o IMC temos que ele é uma variável quantitativa contínua. O objetivo é identificar se há diferenças significativas nos valores de IMC entre esses esportes, bem como entender quais esportes tendem a ter IMC mais altos ou mais baixos.

O cálculo do IMC é dado pela seguinte fórmula:

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

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.

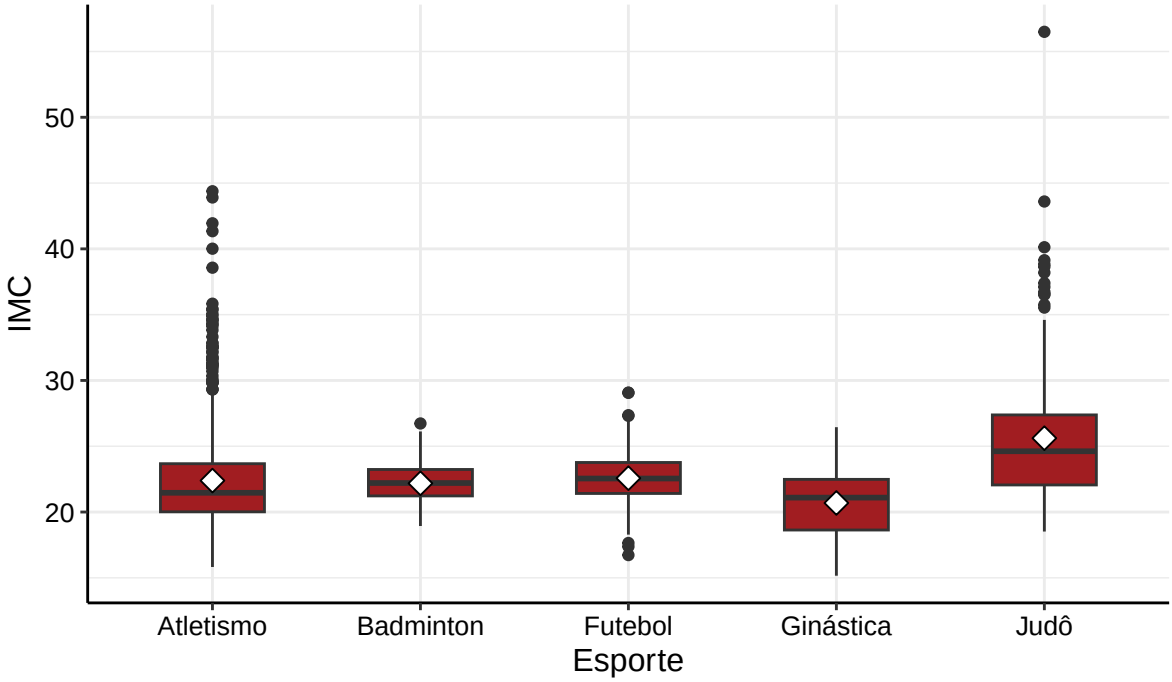


Figura 5: Boxplot da Comparação do IMC entre Esportes

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.

Quadro 1: Estatísticas descritivas por esporte

Esporte	Tamanho Amostral	Média	Mediana	Desvio Padrão	Min	Max
Atletismo	691	22,39	21,46	4,01	15,8	44,4
Badminton	92	22,18	22,20	1,59	18,9	26,7
Futebol	447	22,57	22,55	1,77	16,7	29,1
Ginástica	200	20,69	21,09	2,42	15,2	26,4
Judô	231	25,61	24,62	5,05	18,5	56,5

No Atletismo, a média do IMC é de aproximadamente 22,39, com uma mediana de 21,46 e um desvio padrão de 4,01. Isso indica que, embora a maioria dos atletas de

Atletismo tenha um IMC próximo à média, há certa variabilidade, com valores extremos como o máximo de 44,4, sugerindo diferenças significativas entre os atletas dessa modalidade. Para os atletas de Badminton, a média de IMC é de 22,18 e a mediana é de 22,20, com um desvio padrão de apenas 1,59. Esse baixo desvio padrão indica uma menor variabilidade nos valores de IMC, sugerindo que a maioria dos atletas de Badminton possui valores próximos da média, sem grandes desvios.

No Futebol, a média de IMC é ligeiramente superior, de 22,57, e a mediana é de 22,54, com um desvio padrão de 1,77. Essa baixa variabilidade também indica que a maioria dos jogadores apresenta valores de IMC semelhantes, concentrando-se em torno da média. Na Ginástica, a média do IMC é de 20,69, com uma mediana de 21,09 e um desvio padrão de 2,42. Esse valor relativamente baixo de desvio padrão aponta para uma variabilidade pequena, com a maioria dos ginastas concentrados em torno da média, refletindo uma distribuição homogênea de IMC entre os atletas.

Por fim, no Judô, a média do IMC é a mais alta entre os esportes analisados, com 25,61, enquanto a mediana é de 24,62 e o desvio padrão é de 5,05. Esse alto desvio padrão indica uma grande variabilidade entre os judocas, com um valor máximo de 56,5, sugerindo que alguns atletas possuem IMC significativamente mais altos, o que pode estar relacionado às exigências físicas do esporte, onde uma maior massa corporal é frequentemente vantajosa.

Esses dados mostram que os judocas tendem a ter IMC mais elevados e variáveis, enquanto os ginastas possuem os menores valores de IMC e menor variabilidade. As diferenças de variabilidade e valores médios refletem as demandas físicas específicas de cada esporte, com esportes de força e contato, como o Judô, exigindo maior massa muscular, enquanto esportes que demandam agilidade, como Ginástica e Atletismo, favorecem IMC mais baixos.

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.

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

Quadro 2: P-valor do Teste de Shapiro-Wilk

Teste	P-valor	Decisão do teste
Shapiro-Wilk	<0,001	Rejeita H_0

O valor do p-valor resultante para todos os atletas dos esportes selecionados foi extremamente baixo, 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.

$$\begin{cases} H_0: \text{Não há diferença significativa nas distribuições de IMC entre as modalidades esportivas; ou seja, os valores de IMC têm a mesma mediana entre os grupos.} \\ H_1: \text{Há pelo menos uma diferença significativa nas distribuições de IMC entre as modalidades esportivas; ou seja, os valores de IMC diferem entre os grupos.} \end{cases}$$

Quadro 3: P-valor do Teste de Kruskal-Wallis

Teste	Estatística do teste	P-valor	Decisão do teste
Kruskal-Wallis qui-quadrado	1984,40	<0,001	Rejeita H_0

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 a 2016, além de examinar a quantidade de cada tipo de medalha (ouro, prata e bronze) conquistada por eles. As variáveis utilizadas foram nome dos medalhistas variável qualitativa nominal, número de medalhas variável quantitativa discreta,

tipo de medalha variável qualitativa nominal e quantidade de medalhas por tipo variável quantitativa discreta. 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.

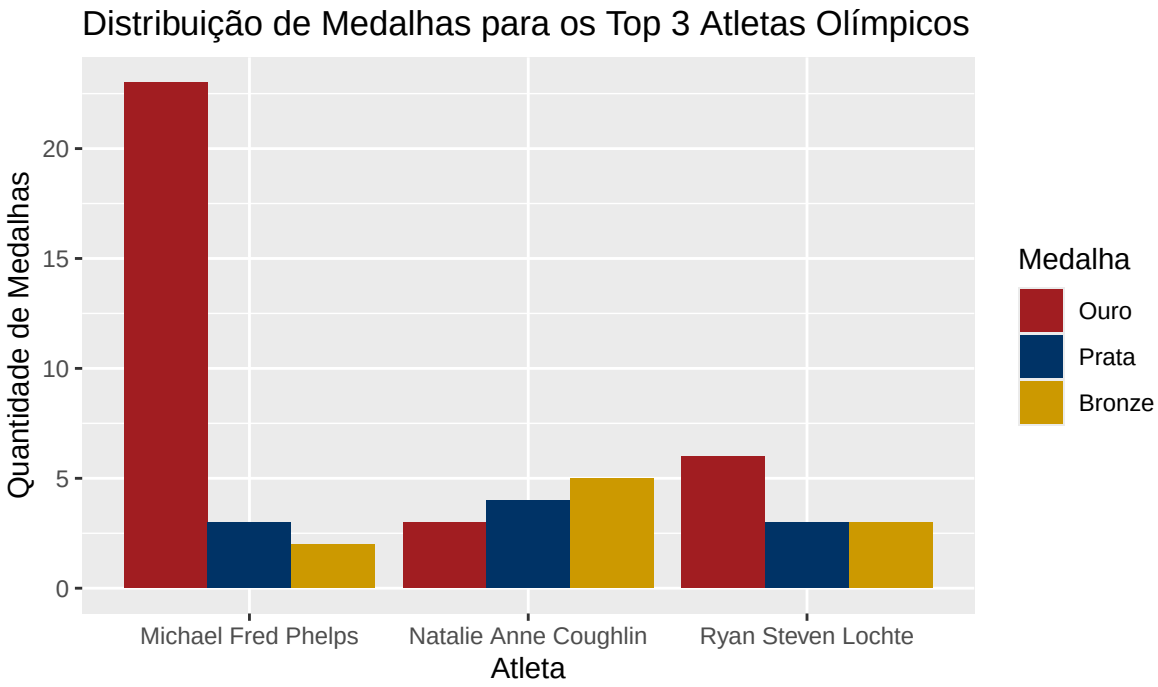


Figura 6: Gráfico de barras da distribuição das medalhas

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.

Quadro 4: Tabela dos Medalhistas e Total de Medalhas

Nome	Total de Medalhas
Michael Fred Phelps	28
Natalie Anne Coughlin	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.

Quadro 5: Tabela da Distribuição das Medalhas

Nome	Medalha	Quantidade
Michael Fred Phelps	Ouro	23
Michael Fred Phelps	Prata	3
Michael Fred Phelps	Bronze	2
Natalie Anne Coughlin	Ouro	3
Natalie Anne Coughlin	Prata	4
Natalie Anne Coughlin	Bronze	5
Ryan Steven Lochte	Ouro	6
Ryan Steven Lochte	Prata	0
Ryan Steven Lochte	Bronze	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.

Podemos confirmar matematicamente fazendo um teste de independência.

$$\begin{cases} H_0 : \text{A diferença entre homens e mulheres segue uma distribuição simétrica em torno de zero} \\ H_1 : \text{A diferença entre homens e mulheres não segue uma distribuição simétrica em torno de zero} \end{cases}$$

Quadro 6: Resultados do teste Qui-Quadrado

Estatística	Estatísticado teste	Graus de Liberdade	P-valor	Decisão do teste
X-quadrado	12.77	4	0,012	Rejeita H_0

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 12,77 com 4 graus de liberdade e um valor-p de 0,012. 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. A altura para essa análise pode variar de 1,39 a 2,07 e o peso para essa análise pode variar de 30,9 a 174,9.

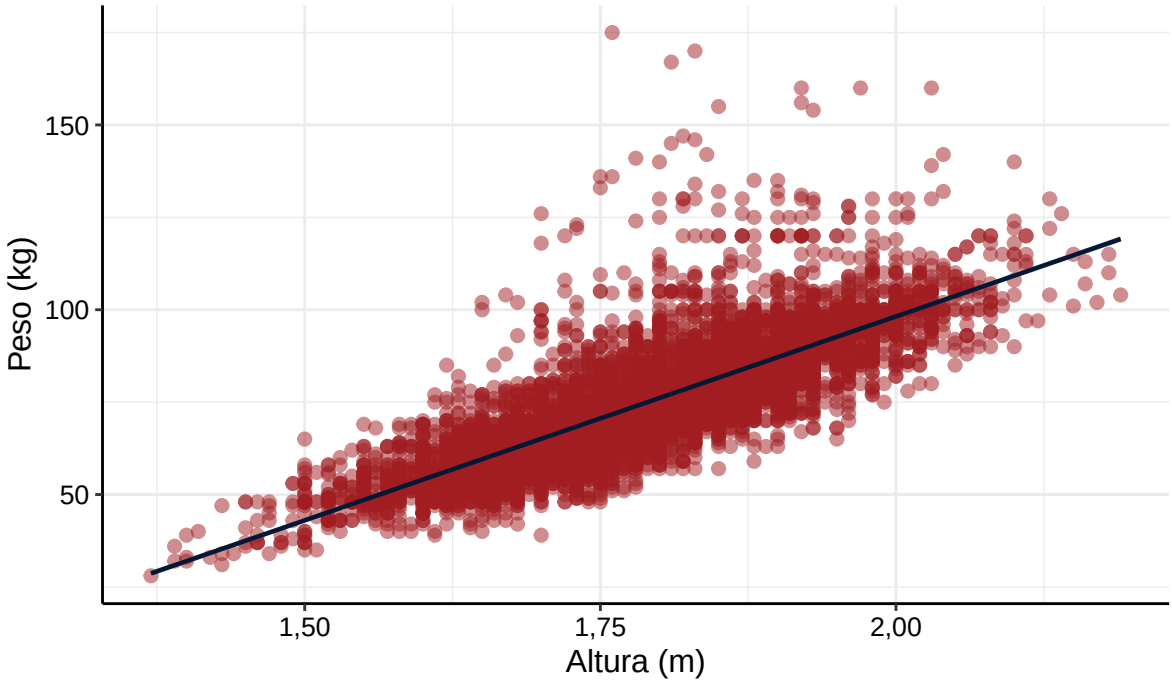


Figura 7: Gráfico de disperssão da Altura e Peso dos Atletas

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.

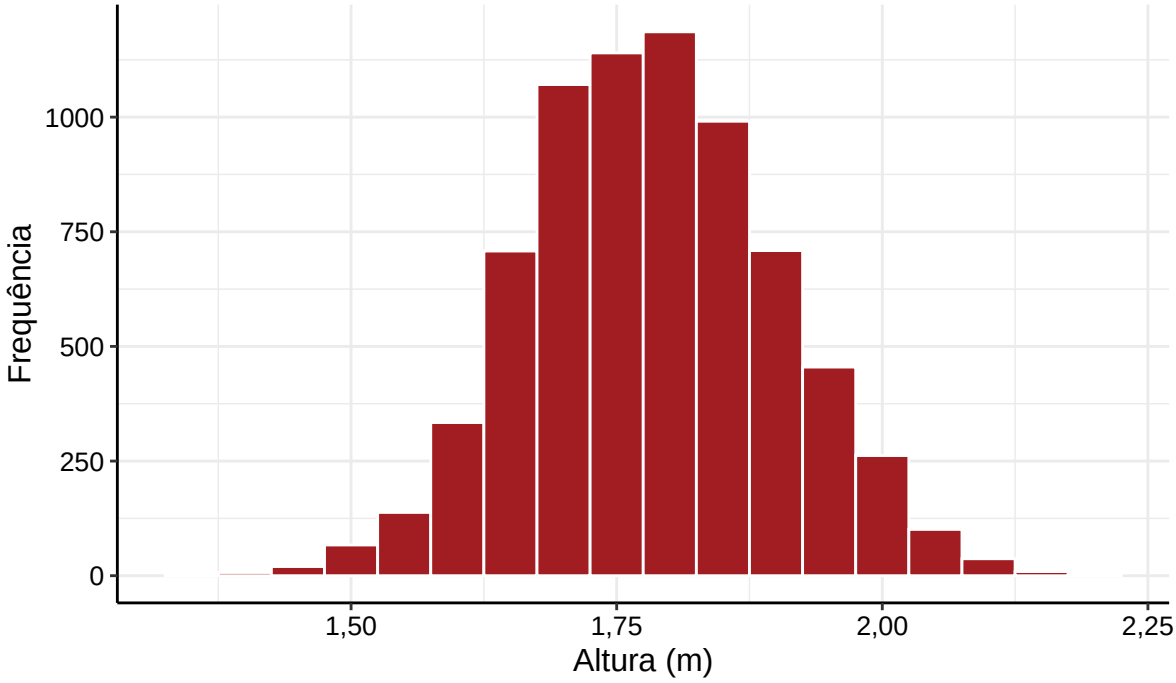


Figura 8: Distribuição da Altura dos Atletas

O gráfico da distribuição da altura dos atletas apresenta uma forma aproximadamente simétrica, centrada em torno de 1,75 metros, sugerindo que a maioria dos atletas tem altura próxima desse valor. A distribuição tem um padrão semelhante ao de uma distribuição normal, embora possa haver uma leve assimetria ou variação nas caudas, o que seria melhor avaliado com testes específicos de normalidade.

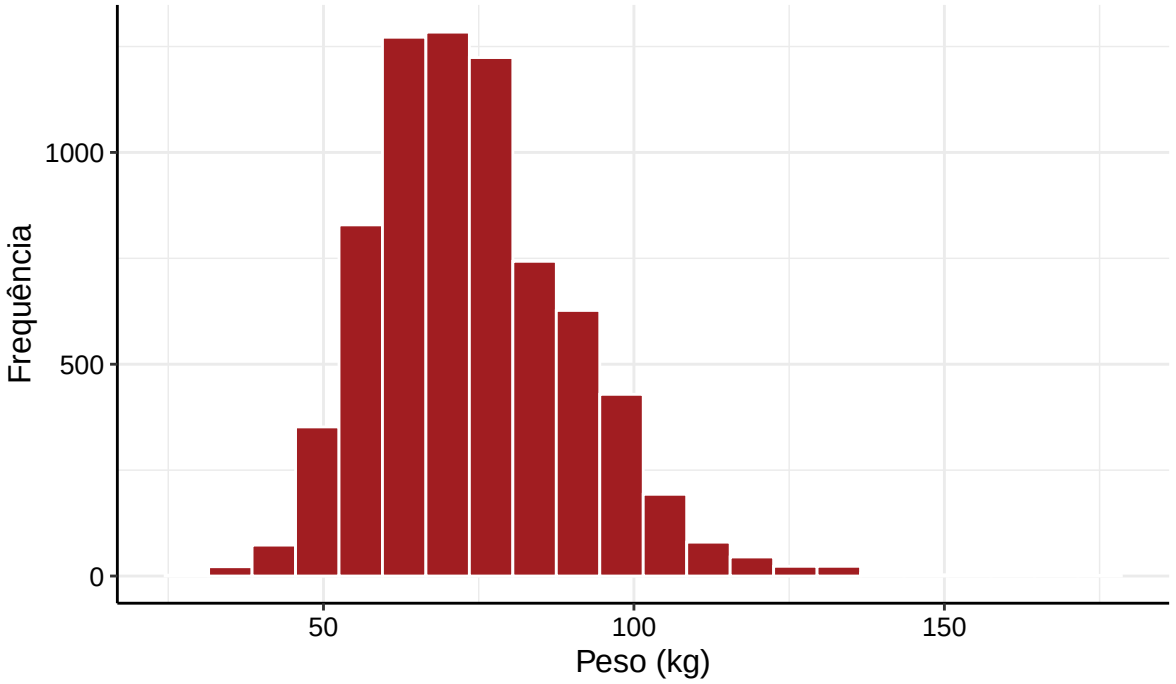
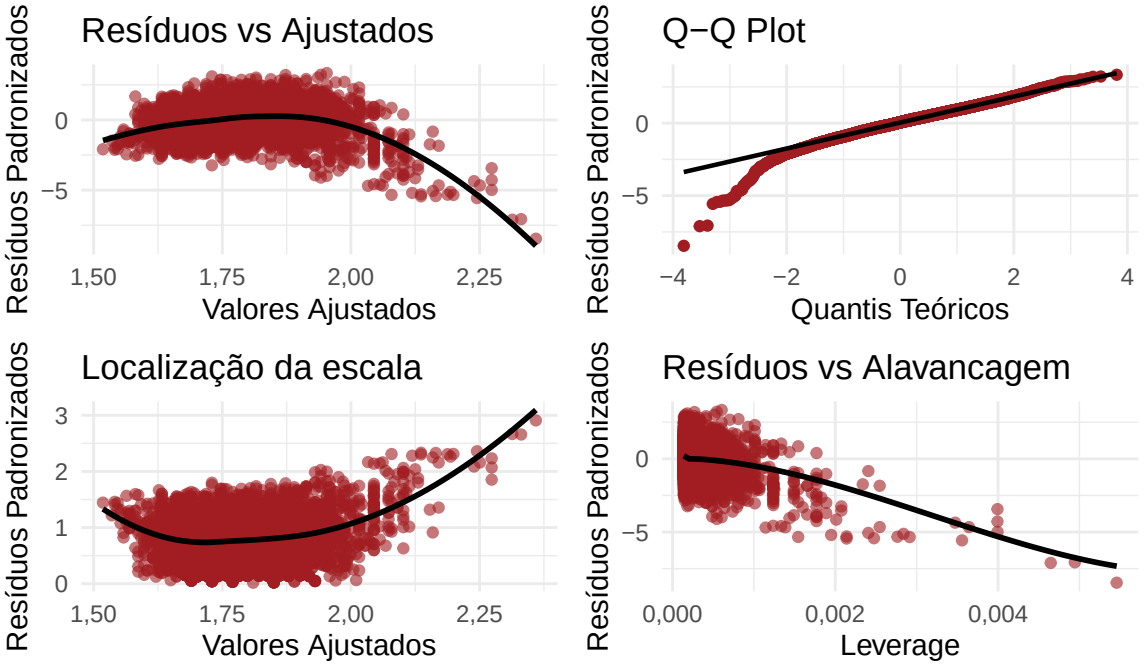


Figura 9: Distribuição do Peso dos Atletas

O gráfico da distribuição do peso dos atletas mostra que é mais assimétrica e apresenta uma cauda longa à direita, indicando que há atletas com pesos mais elevados, mas que representam uma menor proporção da amostra. A maior concentração está entre 50 e 80 kg, o que sugere uma tendência mais centralizada dentro dessa faixa. Essa assimetria positiva é comum em variáveis como peso, onde existem poucos valores muito altos.



Para o gráfico resíduos vs ajustados avalia-se a suposição de linearidade e se a variância dos resíduos é constante. Neste caso, a falta de um padrão claro nos resíduos ao redor de zero indica uma possível adequação, mas a leve tendência curvada sugere alguma possível violação de linearidade. No gráfico q-q plot verifica-se a normalidade dos resíduos. No seu gráfico, a maioria dos pontos segue a linha teórica, mas há desvios nas extremidades, o que indica que os resíduos podem não ser perfeitamente normais (presença de caudas mais pesadas). Já o gráfico localizaçãoda escala também verifica-se a homocedasticidade,ou seja, variância constante dos resíduos. Uma tendência ascendente indica que a variância dos resíduos aumenta com os valores ajustados, sugerindo heterocedasticidade. Por fim, para os resíduos vs alavancagem mostra que a influência de pontos específicos. Observações fora das linhas de Cook's distance podem ser pontos influentes que afetam o modelo de forma desproporcional. Neste gráfico, alguns pontos aparecem próximos às linhas, o que indica que esses pontos podem ter grande influência na regressão.

Quadro 7: Medidas resumo da Altura

Estatística	Valor
Média	1,78
Desvio Padrão	0,12
Variância	0,01
Mínimo	1,37
1º Quartil	1,70
Mediana	1,78
3º Quartil	1,86
Máximo	2,19

Com base no quadro 7 temos qua a média da altura é de 1,78 metros, indicando que, em geral, os indivíduos têm uma estatura próxima a esse valor. O desvio padrão de 0,12 metros mostra que há uma variação pequena ao redor da média, sugerindo que a maioria dos indivíduos tem alturas próximas a essa média. A variância, outra medida de dispersão, é de 0,01, o que confirma essa baixa variabilidade. Quanto aos valores extremos e à distribuição da altura, o mínimo é de 1,37 metros e o máximo é de 2,19 metros, o que estabelece o intervalo total das alturas observadas.

Quadro 8: Medidas resumo do Peso

Estatística	Valor
Média	74,15
Desvio Padrão	16,25
Variância	264,13
Mínimo	28,00
1º Quartil	63,00
Mediana	72,00
3º Quartil	84,00
Máximo	175,00

De acordo com o quadro 8 a média do peso dos atletas é de 74,15 kg, indicando o valor central dessa distribuição. O desvio padrão de 16,25 kg indica a variabilidade ou dispersão dos valores em torno da média, sendo assim o peso dos atletas está relativamente disperso. A variância é de 264,13, mopstrando uma alta dispersão dos dados. O mínimo registrado é 28 kg, enquanto o máximo chega a 175 kg, o que mostra uma grande variabilidade de pesos entre os atletas. Os quartis dividem os dados em quatro partes iguais, sendo o primeiro quartil 63 kg, a mediana 72 kg e o terceiro quartil 84 kg. A mediana indica que metade dos atletas tem peso abaixo de 72 kg e a outra metade acima, enquanto o intervalo interquartil (84 - 63 = 21) sugere que a maior parte dos atletas tem pesos concentrados entre 63 kg e 84 kg.

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

Quadro 9: Tabela da Normalidade da Altura dos Atletas

Estatística	Valor	P-valor	Decisão do teste
A	6,42	<0,001	Rejeita H_0

Quadro 10: Tabela da Normalidade do Peso dos Atletas

Estatística	Valor	P-valor	Decisão do teste
A	40,83	<0,001	Rejeita H_0

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.

Quadro 11: Resultados da Regressão Linear Simples para Altura em função do Peso

Variável	Estimativa	Erro Padrão	Estatística t	P-valor
Intercepto	1,36	0,01	348,08	<0,001
Peso	0,01	0,001	111,37	<0,001

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.

Quadro 12: Resultado do Teste F da Regressão

Teste	P-valor
Teste F da Regressão	<0,001

O teste F realizado na regressão revelou um p-valor de < 0,001, indicando que o efeito do peso sobre a altura é estatisticamente significativo ao nível de 5%. Isso implica que a variação no peso dos atletas está significativamente associada a variações na altura, com uma probabilidade muito baixa de que essa associação seja devida ao acaso. Esse resultado nos permite concluir que há uma relação linear significativa entre as variáveis peso e altura para o conjunto de dados analisado. Contudo, é importante destacar que, embora a significância estatística indique uma associação, ela não revela a intensidade da relação, que pode ser observada pelo coeficiente de correlação de Pearson (0,79), além do coeficiente de determinação na análise de regressão.

4 Conclusões

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.

Notou-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.

Observou-se que os três principais medalhistas olímpicos – Phelps, Coughlin e Lochte – têm padrões distintos de conquistas de medalhas. Phelps se destacou pela grande quantidade de ouros, enquanto Coughlin e Lochte apresentaram um equilíbrio maior entre as diferentes medalhas. Esse padrão foi confirmado pelo teste qui-quadrado, evidenciando a diversidade no desempenho dos atletas e como diferentes fatores podem influenciar o sucesso olímpico.

Por fim, confirmou-se uma correlação positiva significativa entre o peso e a altura dos atletas, reforçando a ideia de que, em muitas modalidades, existe uma relação entre essas características físicas. Esse resultado abre portas para investigações mais detalhadas sobre como as características corporais podem impactar o desempenho nos Jogos Olímpicos.