Análise do desempenho dos atletas de elite

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

Requerente:
ESTAT

Brasília, 12 de novembro de 2024.

ESTAT
consultoria estatística

Sumário

1 Introdução

O seguinte projeto tem como objetivo realizar uma análise estatística do desempenho dos atletas de elite da House of Excellence que participaram das Olimpíadas entre 2000 e 2016. As análises incluem estatísticas descritivas e comparações entre grupos, visando entender fatores como a performance das medalhistas mulheres, a variação do Índice de Massa Corporal (IMC) por esporte, o número de medalhas conquistadas por atletas destacados e a relação entre peso e altura.

O banco de dados utilizado foi coletado pelo cliente, contendo informações sobre atletas e suas performances nas Olimpíadas, incluindo as variáveis nome, gênero, idade, altura, peso, país, esporte, evento, tipo de medalha e ano de conquista da medalha. A amostra inclui dados de diferentes países e esportes, proporcionando uma visão abrangente e probabilística do desempenho atlético. Essas informações são cruciais para as análises, permitindo uma compreensão mais profunda dos fatores que impactam o desempenho.

Para a realização das análises, foi utilizado o software R, versão 4.2.0. Este software é amplamente reconhecido na comunidade estatística por suas capacidades de manipulação e visualização de dados, além de fornecer uma ampla gama de pacotes para análises estatísticas avançadas. A utilização do R garantiu a precisão e a eficácia na execução das análises propostas, assim como na geração de gráficos e relatórios visuais que complementam os resultados obtidos.

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 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.

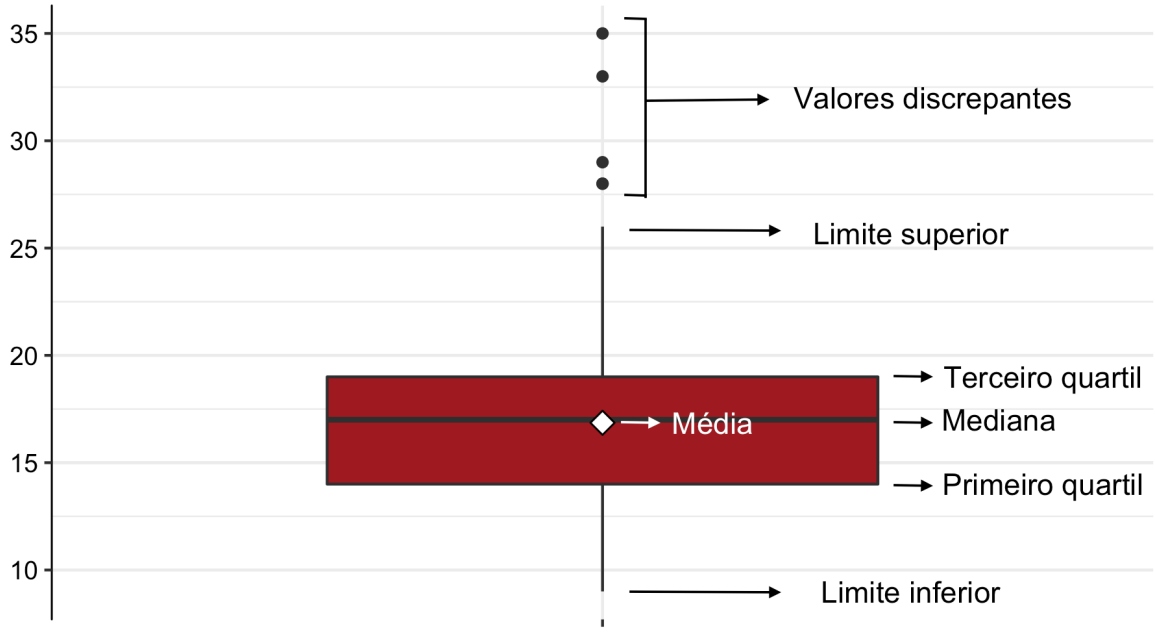

Figura 1: Exemplo de boxplot

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.

2.8 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.

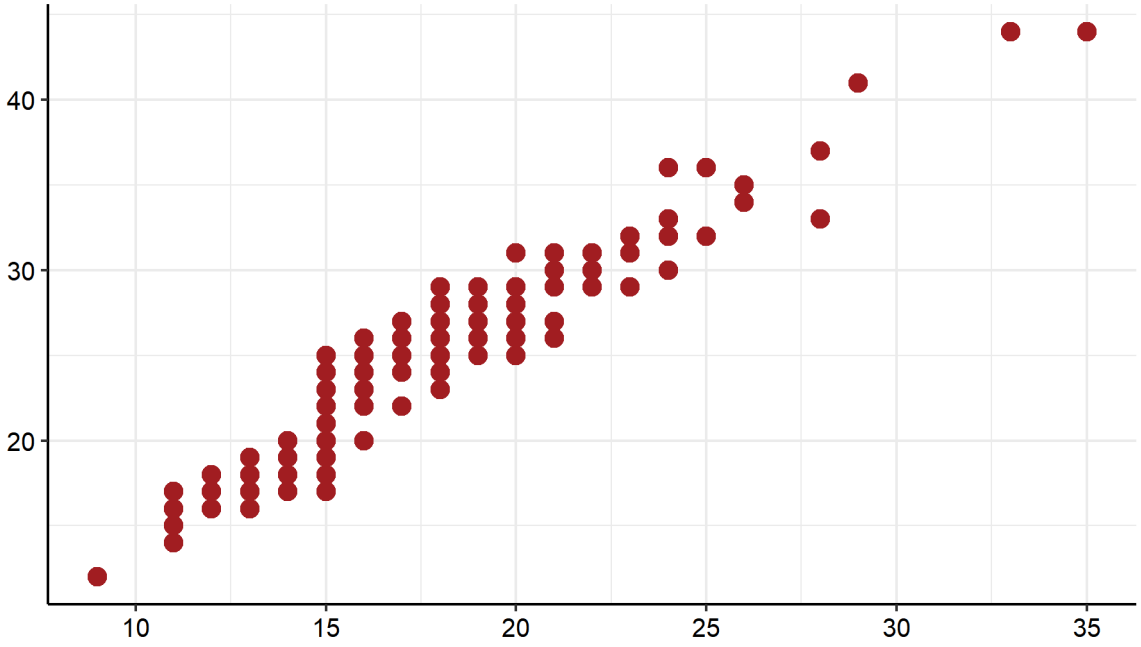

Figura 2: Exemplo de Gráfico de Dispersão

2.9 Tipos de Variáveis

2.9.1 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)

2.9.2 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)

2.10 Definição para Testes

2.10.1 Teste de Hipóteses

O teste de hipóteses tem como objetivo fornecer uma metodologia para verificar se os dados das amostras possuem indicativos que comprovem, ou não, uma hipótese previamente formulada. Ele é composto por duas hipóteses:

$$\begin{cases} H_0 : \text{hipótese a ser testada (chamada de hipótese nula)} \\ H_1 : \text{hipótese alternativa que será aceita caso a hipótese nula seja rejeitada} \end{cases}$$

Essa decisão é tomada por meio da construção de uma região crítica, ou seja, região de rejeição do teste.

2.10.2 Tipos de teste: bilateral e unilateral

Para a formulação de um teste, deve-se definir as hipóteses de interesse. Em geral, a hipótese nula é composta por uma igualdade (por exemplo, $H_0 : \theta = \theta_0$). Já a hipótese alternativa depende do grau de conhecimento que se tem do problema em estudo. Assim, tem-se três formas de elaborar H_1 que classificam os testes em duas categorias:

- **Teste Bilateral**:
  Esse é o teste mais geral, em que a hipótese alternativa consiste em verificar se existe diferença entre os parâmetros de interesse, independentemente de um ser maior ou menor que o outro. Dessa forma, tem-se:

$$H_1 : \theta \neq \theta_0$$

- **Teste Unilateral**:
  Dependendo das informações que o pesquisador possui a respeito do problema e os questionamentos que possui, a hipótese alternativa pode ser feita de forma a verificar se existe diferença entre os parâmetros em um dos sentidos. Ou seja:

$$H_1 : \theta < \theta_0$$

ou

$$H_1 : \theta > \theta_0$$

2.10.3 Tipos de Erros

Ao realizar um teste de hipóteses, existem dois erros associados: Erro do Tipo I e Erro do Tipo II.

- **Erro do Tipo I**:
  Esse erro é caracterizado por rejeitar a hipótese nula (H_0) quando essa é verdadeira. A probabilidade associada a esse erro é denotada por α, também conhecido como nível de significância do teste.

- **Erro do Tipo II**:
  Ao não rejeitar H_0 quando, na verdade, é falsa, está sendo cometido o Erro do Tipo II. A probabilidade de se cometer este erro é denotada por β.

2.10.4 Nível de significância (α)

O nível de significância do teste é o nome dado à probabilidade de se rejeitar a hipótese nula quando essa é verdadeira; essa rejeição é chamada de erro do tipo I. O valor de α é fixado antes da extração da amostra e, usualmente, assume 5%, 1% ou 0,1%.

Por exemplo, um nível de significância de $\alpha = 0,05$ (5%) significa que, se for tomada uma grande quantidade de amostras, em 5% delas a hipótese nula será rejeitada quando não havia evidências para essa rejeição, isto é, a probabilidade de se tomar a decisão correta é de 95%.

2.10.5 Estatística do Teste

A estatística do teste é o estimador que será utilizado para testar se a hipótese nula (H_0) é verdadeira ou não. Ela é escolhida por meio das teorias estatísticas.

2.10.6 P-valor

O P-valor, ou nível descritivo, é uma medida utilizada para sintetizar o resultado de um teste de hipóteses. Ele também pode ser chamado de *probabilidade de significância* do teste e indica a probabilidade de se obter um resultado da estatística de teste mais extremo do que o observado na presente amostra, considerando que a hipótese nula é verdadeira. Dessa forma, rejeita-se H_0 quando P-valor $< \alpha$, porque a chance de uma nova amostra possuir valores tão extremos quanto o encontrado é baixa, ou seja, há evidências para a rejeição da hipótese nula.

2.10.7 Intervalo de Confiança

Quando calcula-se um estimador pontual para o parâmetro, não é possível definir qual a possível magnitude do erro que se está cometendo. Com o objetivo de associar um erro à estimativa, são construídos os intervalos de confiança que se baseiam na distribuição amostral do estimador pontual.

Dessa forma, considere T um estimador pontual para θ e que a distribuição amostral de T é conhecida. O intervalo de confiança para o parâmetro θ será dado por t_1 e t_2, tal que:

$$P(t_1 < \theta < t_2) = \gamma$$

A probabilidade γ é estabelecida no início do estudo e representa o nível de confiança do intervalo. A interpretação desse resultado é que, se forem tiradas várias amostras de mesmo tamanho e forem calculados intervalos de confiança para cada uma, $100 \times \gamma\%$ dos intervalos irão conter o parâmetro θ. Assim, ao calcular um intervalo, pode-se dizer que há $100 \times \gamma\%$ de confiança de que o intervalo contém o parâmetro de interesse.

2.10.8 Teste de Normalidade

Os testes de normalidade são utilizados para verificar se uma variável aleatória segue um distribuição Normal de probabilidade ou não. Eles são muito importantes, pois impactam em qual teste deve ser utilizado em uma análise futura. Se o resultado do teste confirmar que a variável segue uma distribuição normal, procedimentos paramétricos podem e devem ser utilizados. Caso contrário, os métodos não paramétricos são mais recomendados.

2.10.8.1 Teste de Normalidade de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é usado para determinar se duas distribuições de probabilidade diferem uma da outra. É baseado na diferença entre a função de distribuição acumulada teórica $F_0(x)$ e a função de distribuição acumulada da amostra

$S_n(x)$. A função $S_n(x)$ é definida como a proporção das observações da amostra que são menores ou iguais a x.

O teste possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue o modelo proposto} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

Se a hipótese nula é verdadeira, espera-se que as diferenças entre $F_0(x)$ e $S_n(x)$ sejam pequenas e estejam dentro dos limites dos erros aleatórios. O teste de Kolmogorov-Smirnov focaliza a maior dessas diferenças. No caso do teste de normalidade de Kolmogorov-Smirnov, a função de distribuição acumulada teórica $F_0(x)$ é a função de distribuição acumulada da normal, com média e variância estimadas pela amostra. Este teste é mais recomendado para amostras grandes sem *outliers*.

2.10.9 Teste de Comparação de Médias

2.10.9.1 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.

2.10.10 Teste de Independência

2.10.10.1 Testes Qui-Quadrado

Os testes a seguir utilizam como base a estatística χ^2, apresentando mudanças nos graus de liberdade da sua distribuição de acordo com o teste que será utilizado. No geral,

$$\chi^2_v = \sum \frac{(o_i - e_i)^2}{e_i}$$

em que v expressa os graus de liberdade, o_i é a frequência observada e e_i é chamado de valor esperado e representa a frequência que seria observada se H_0 fosse verdadeira.

Esse teste tem como objetivo verificar se existe associação entre duas variáveis, sendo mais recomendado para variáveis qualitativas (principalmente nominais). O princípio básico deste método é comparar proporções, ou seja, as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Para esse teste, as hipóteses podem ser escritas como:

$$\begin{cases} H_0 : \text{A variável X é independente da variável Y} \\ H_1 : \text{A variável X depende da variável Y} \end{cases}$$

Este teste é baseado no cálculo dos valores esperados. Os valores esperados são os valores que seriam observados caso a hipótese nula fosse verdadeira:

$$e_{ij} = \frac{(total\ da\ linha\ i) \times (total\ da\ coluna\ j)}{total\ geral}$$

Para isso, utiliza-se a seguinte estatística:

$$\chi^2_v = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

- $e_{ij} =$ valor esperado na i-ésima linha e na j-ésima coluna
- $o_{ij} =$ valor observado na i-ésima linha e na j-ésima coluna
- $v = (r-1)(s-1)$ representa o número de graus de liberdade
- $r =$ número total de linhas
- $s =$ número total de colunas

Então, sob a hipótese de H_0 ser verdadeira, a estatística do teste seguirá a distribuição χ^2_v.

Para que a aproximação Qui-Quadrado seja satisfatória, é preciso que a amostra seja relativamente grande, com todos os valores esperados maiores ou iguais a 5 ou no máximo 20% deles seja menor que 5 com todos maiores que 1. Caso isso não ocorra, utiliza-se a correção de Yates.

2.10.11 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.

3 Análises

3.1 Top 5 países com maior número de mulheres medalhistas

O objetivo dessa análise é ver quais países possuem o maior número de mulheres que conquistaram medalhas, em ordem do país com maior número de mulheres medalhistas para o país com menor número.

Para isso, foi criado um gráfico de barras verticais, ideal para analisar uma variável quantitativa discreta, que são valores enumeráveis como a quantidade de medalhas e uma variável qualitativa nominal que são os nomes dos países. No gráfico, os países estão representados no eixo X em ordem do maior número de mulheres medalhistas para o menor e a quantidade de medalhas no eixo Y, quantificados por frequência relativa que é utilizada para a comparação entre classes de categorias, como os países.

Para melhor vizualização, acompanhe o gráfico abaixo:

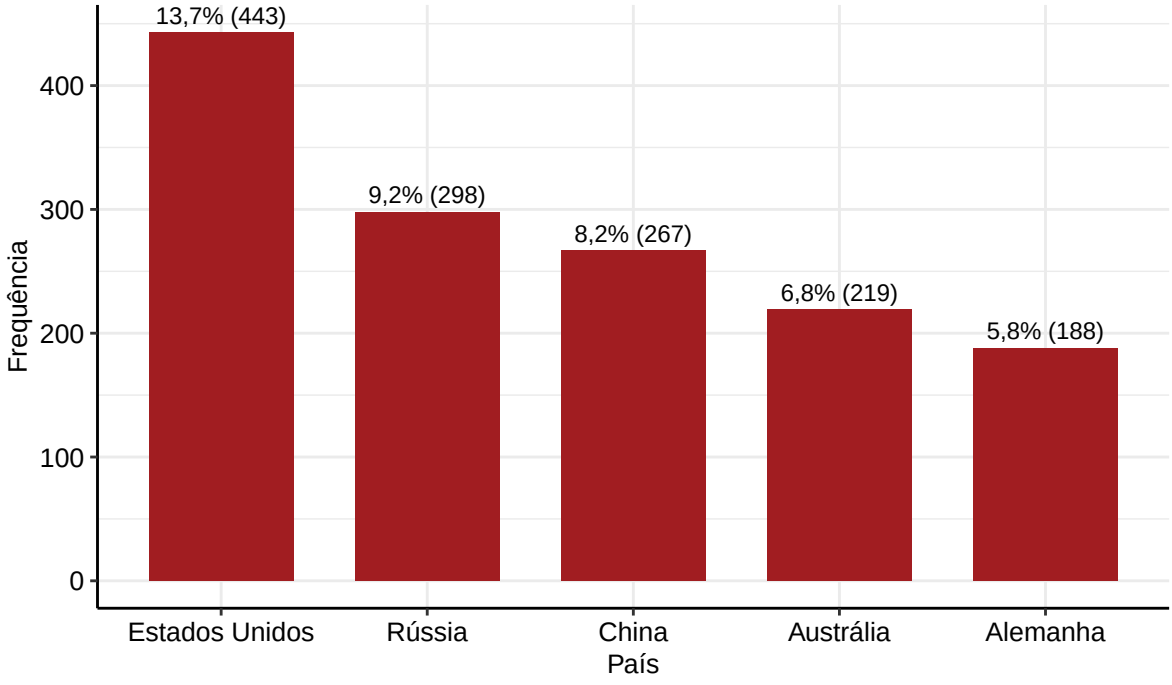

Figura 3: Gráfico de colunas do total de medalhas de mulheres medalhistas

É possível vizualizar a partir da **Figura 3** que em primeiro lugar está os Estados Unidos, com o total de 433 medalhistas, seguido da Rússia com 298 medalhistas, seguido da China com 267 medalhistas, seguida da Austrália com 219 medalhistas, seguida da Alemanha com 188 medalhistas.

É possível ver uma diferença entre os 5 primeiros países, sendo os Estados Unidos com 145 medalhas a mais que o segundo colocado e 255 medalhas a mais que o quinto colocado, diferença essa maior que qualquer total de medalhas de outro país.

Assim, os Estados Unidos tem mais medalhas que a soma do total de medalhas que a Austrália e Alemanha juntas. Dessa forma, conclui-se que os Estados Unidos é um país que prepara com excelência as atletas para as Olimpíadas.

Além disso, é possível ver que não há diferença entre a Rússia, China e Austrália quanto o total de mulheres medalhistas, visto que sua diferença entre as medalhas é baixa. Dessa forma, esses países se equivalem na preparação de atletas medalhistas.

Essa análise não apenas destaca a importância do investimento em esportes femininos, como enfatiza países com estratégias eficazes que podem ser implementadas para fomentar o sucesso de mulheres atletas em nível internacional. Portanto, enquanto os Estados Unidos estabelecem um padrão elevado, a competitividade entre Rússia, China e Austrália também ressalta a crescente importância do esporte feminino globalmente.

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

O objetivo dessa análise é observar o IMC dos esportes selecionados, estes sendo, ginástica, futebol, judô, atletismo e badminton. O IMC é uma variável quantitativa contínua, ou seja apresenta valores resultados de medições, permitindo construir um boxplot para perceber como os dados estão distribuídos. Para ver se há uma diferença nos valores de IMC entre os diferentes esportes, foram usados os testes de hipóteses de Kolmogorov-Smirnov e Kruskal-Wallis

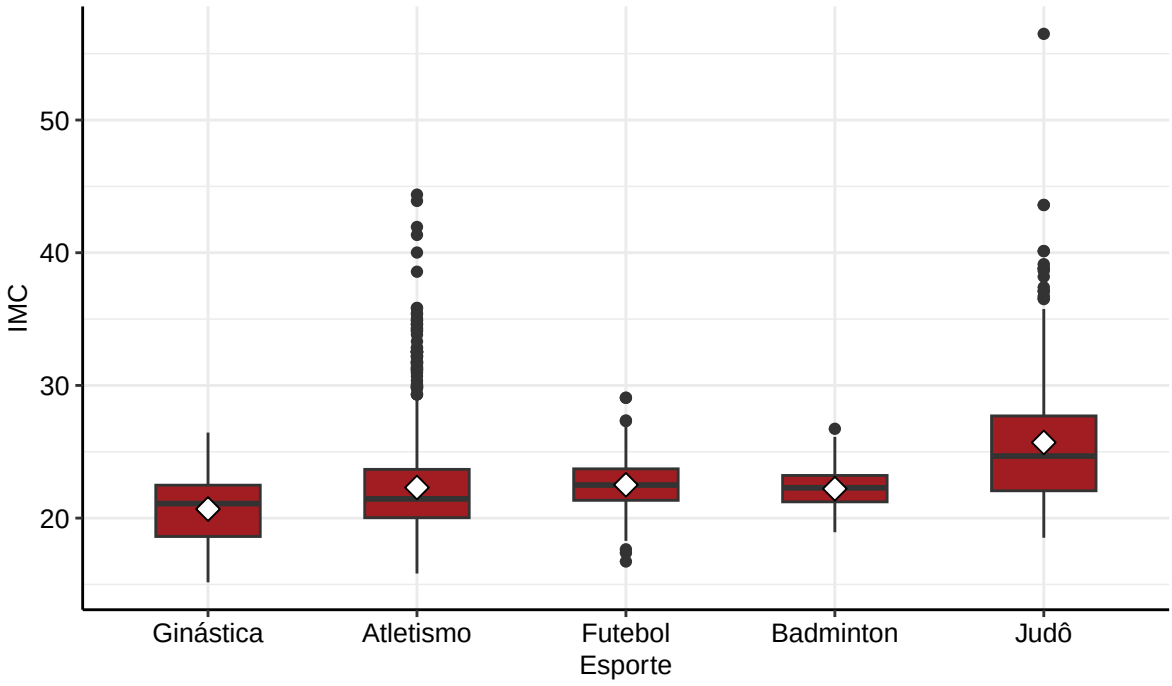

Figura 4: Boxplot do IMC

Quadro 1: Medidas resumo do IMC

Estatística	Ginástica	Atletismo	Futebol	Badminton	Judô
Média	20,68	22,30	22,51	22,21	25,70
Desvio Padrão	2,38	3,86	1,73	1,50	5,12
Variância	5,67	14,92	2,99	2,26	26,23
Mínimo	15,16	15,82	16,73	18,94	18,52
1º Quartil	18,61	20,03	21,34	21,22	22,06
Mediana	21,09	21,45	22,49	22,28	24,68
3º Quartil	22,48	23,67	23,71	23,21	27,70
Máximo	26,45	44,38	29,07	26,73	56,50

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

Quadro 2: P-valor do Teste de Kolmogorov-Smirnov do IMC

Variável	P-valor	Decisão do teste
IMC	<0,010	Rejeita H_0

$$\begin{cases} H_0 : \text{Não há diferença significativa nos valores de IMC entre os diferentes esportes} \\ H_1 : \text{Há diferença significativa nos valores de IMC entre os diferentes esportes} \end{cases}$$

Quadro 3: P-valor do Teste de Kruskal-Wallis da média do IMC

Variável	P-valor	Decisão do teste
IMC	<0,010	Rejeita H_0

Verifica-se a partir do **Figura 4** que as medianas de Ginástica e Atletismo são próximas, assim como as de Futebol e Badminton, já a de Judô está acima dos terceiros quartis dos outros esportes, mostrando que o valor que está em 50% para o IMC de Judô está acima de 75% dos outros esportes, revelando que os dados de IMC para os judocas estão muito elevados. Além disso, a amplitude interquartil, que indica a dispersão dos dados, é pequena para Futebol e Badminton, intemediária para Ginástica e Atletismo e alta para judô, revelando a variabilidade do índice IMC nesses esportes. Quanto aos outliers, valores atípicos de uma categoria, estão

evidentes no Atletismo e Judô, revelando que há muitos atletas nesses esportes que apresentam alto IMC, além do comum. Quanto à simetria dos dados, tem-se que é simétrica para o Futebol, Badminton e Judô, revelando que os dados estão simetricamente distribuidos, ou seja, não tendam para para os extremos e assimétrica à direita para o Atletismo, dados próximos de um IMC mais baixo e assimétrica à esquerda para a Ginástica, revelando que os dados tendêm para um IMC maior entre essa categoria. Comparando os dados entre os esportes, vê-se claramente que os IMC menores pertencem aos atletas da Ginástica, intemediários para os atletas de Atletismo, Futebol e Badminton e IMC maior para os judocas.

O resultado do teste de normalidade Shapiro-Wilk com 5% de nível de significância revelou que os dados não estão em conformidade com a distribuição normal, pois o p_valor foi extremamente baixo, próximo a 0. Assim, para verificar se as médias do IMC dos esportes são iguais, foi realizado o teste de Kruskal-Wallis, que indica que há uma diferença significativa nos valores de IMC entre os diferentes esportes, com p-valor extremamente baixo, o que leva a rejeitar a hipótese nula com 5% de nível de significância.

Esses resultados sugerem a necessidade de uma avaliação cuidadosa das exigências físicas de cada esporte e a importância de estratégias nutricionais e de treinamento adequadas para promover a saúde e o desempenho atlético. A análise do IMC não apenas fornece insights sobre a condição física dos atletas, mas também pode informar decisões sobre intervenções que visem melhorar a saúde e o bem-estar dentro dessas modalidades esportivas.

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

O objetivo dessa análise é entender quais são os 3 medalhistas com maior número de medalhas no total, observar entre eles a quantidade de cada tipo de medalha que cada um conquistou e se existe relação entre o medalhista e cada tipo de medalha conquistada. Para isso, foi criado um gráfico de barras para vizualizar a proporção do tipo de medalhas no total de medalhas de cada atleta, sendo a variável “tipo de medalha” qualitativa ordinal e “atleta” qualitativa nominal.

Para avaliar se há relação entre o medalhista e cada tipo de medalha conquistada, ou seja, se o medalhista e o tipo de medalha são variáveis independentes, usou-se o teste de hipótese de Qui Quadrado de independência utilizado quando se quer testar a independência de variáveis com poucos dados de observação.

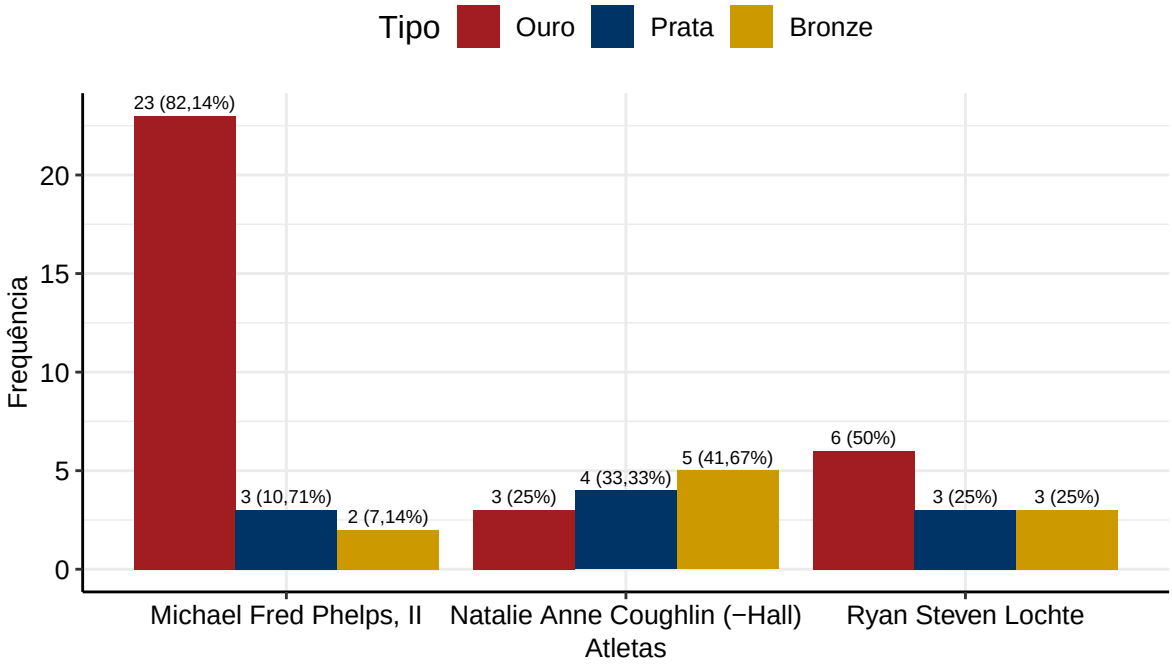

Figura 5: Gráfico de colunas do tipo de medalha

$$\begin{cases} H_0 : \text{Não há relação entre o medalhista e o tipo de medalha (são independentes)} \\ H_1 : \text{Há relação entre o medalhista e o tipo de medalha (não são independentes)} \end{cases}$$

Quadro 4: P-valor do teste Qui-Quadrado de Independência entre o medalhista e o tipo de medalha

Variável	P-valor	Decisão do teste
Medalhas	0,012	Rejeita H_0

A análise estatística revela informações revelantes sobre a distribuição de medalhas entre os atletas. Por meio da **Figura 5**, vemos que Phelps acumula um total de 28 medalhas, das quais 23 são de ouro, 3 de prata e 2 de bronze. Este desempenho não apenas o posiciona como o atleta com o maior número absoluto de medalhas, mas também evidencia uma predominância acentuada nas medalhas de ouro, que representam aproximadamente 82% de suas conquistas totais.

Em contraste, Natalie Coughlin e Ryan Lochte, ambos com 12 medalhas, apresentam perfis de medalhas diferentes. Coughlin possui uma distribuição de 5 medalhas de bronze, 4 de prata e 3 de ouro, enquanto Lochte tem 3 de bronze, 3 de prata e 6 de ouro. Essa variação na distribuição das medalhas sugere que, embora ambos atletas tenham número total equivalente de medalhas, a qualidade das

conquistas, refletida pela quantidade de medalhas de ouro, é inferior em comparação com Phelps.

A discrepância observada nas medalhas de ouro de Phelps em relação aos outros atletas é notável, pois a soma das medalhas de ouro de Coughlin e Lochte (9) ainda é inferior ao total conquistado por Phelps. Essa diferença ressalta a importância do tipo de medalha na avaliação do desempenho dos atletas, indicando que a medalha de ouro não apenas representa a conquista máxima em competições, como é um indicador crítico de sucesso em comparação com medalhas de menor valor, como as de prata e bronze.

Para investigar a relação entre o tipo de medalha conquistada e o atleta, foi realizado um teste Qui-Quadrado de Independência.Com um nível de significância de 5%, os resultados indicaram que existe uma associação entre o medalhista e o tipo de medalha conquistada, com p-valor igual a 0,012. Isso implica que as variáveis “atleta” e “tipo de medalha” não são independentes, sugerindo que o desempenho em termos de medalhas varia de maneira entre os atletas. Essa conclusão é relevante para a compreensão das dinâmicas competitivas no contexto do esporte, pois indica que fatores como habilidade, treinamento e experiência podem influenciar a capacidade de um atleta em conquistar medalhas de diferentes categorias.

3.4 Variação Peso por Altura

O objetivo dessa análise é entender a relação entre o peso e altura dos atletas para compreender se à medida que o peso aumenta a altura também aumenta, diminui ou não tem diferença entre essas variáveis. Para isso, foi criado um gráfico de dispersão para ilustrar o comportamento conjunto de duas variáveis quantitativas contínuas, onde cada ponto representa uma observação do banco de dados.

Também foram criados gráficos boxplot para cada variável para ver sua distribuição, além do quadro de suas medidas resumo. Para avaliar se há relação entre o peso e a altura dos atletas, usou-se primeiramente o teste de Kolmogorov-Smirnov para verificar se as variáveis seguem uma distribuição normal e depois o teste de correlação de Pearson para ver se há correlação entre essas variáveis.

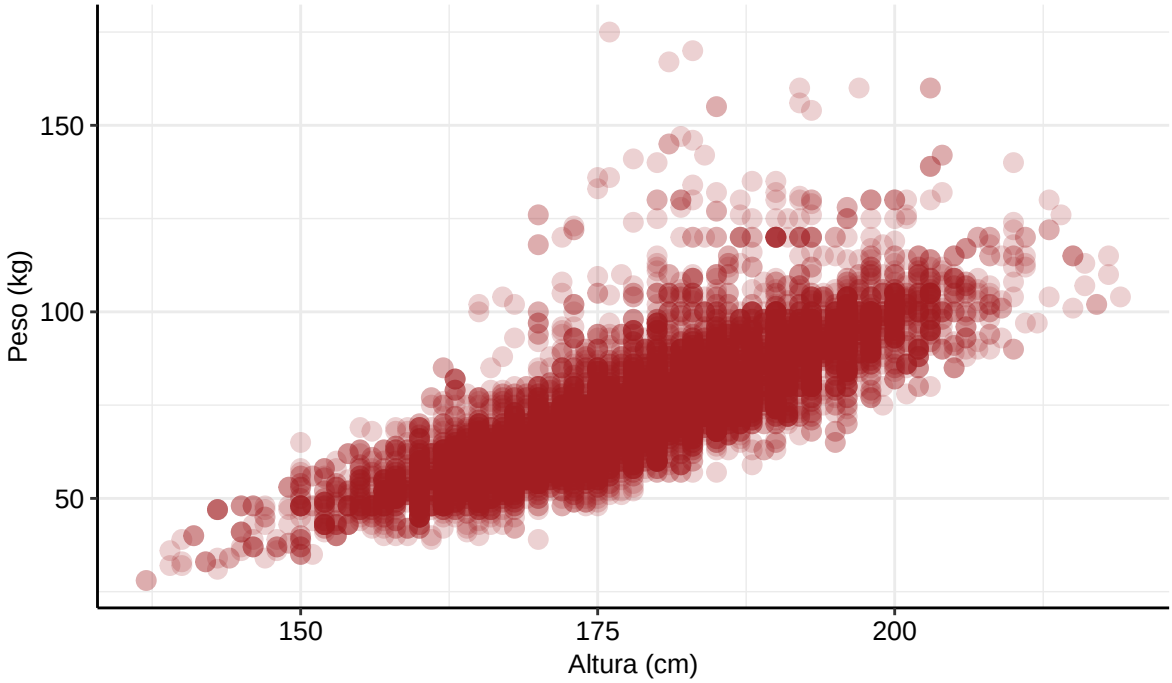

Figura 6: Gráfico de dispersão entre altura e peso

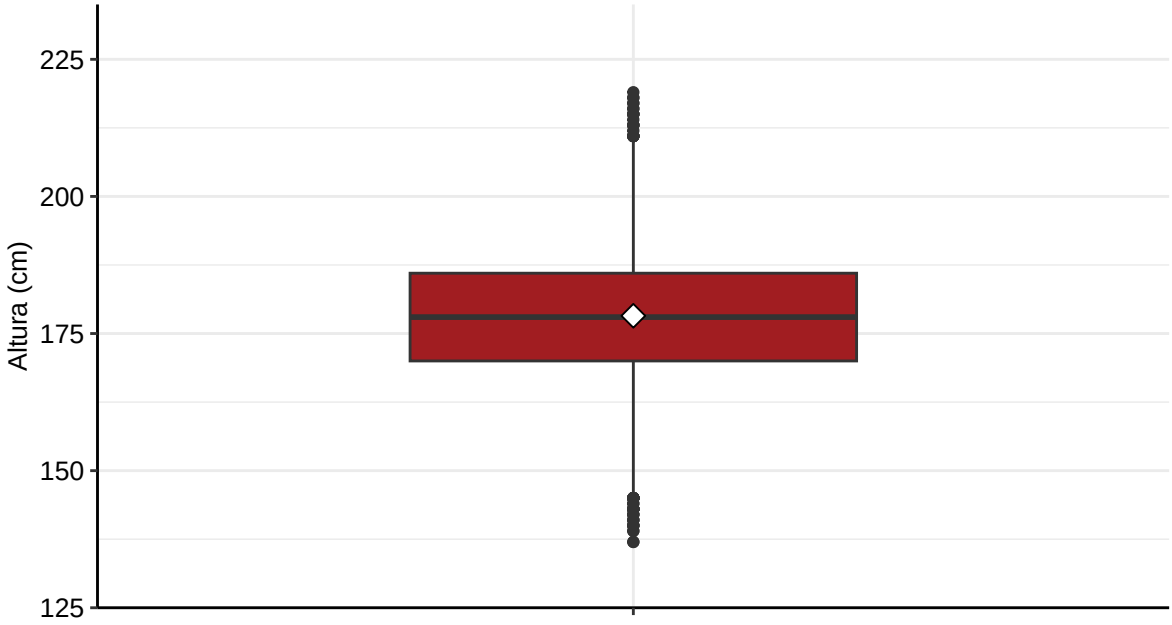

Figura 7: Gráfico boxplot da altura

Quadro 5: Medidas resumo da altura dos atletas

Estatística	Valor
Média	176,11
Desvio Padrão	11,46
Variância	131,24
Mínimo	133,00
1º Quartil	168,00
Mediana	176,00
3º Quartil	184,00
Máximo	226,00

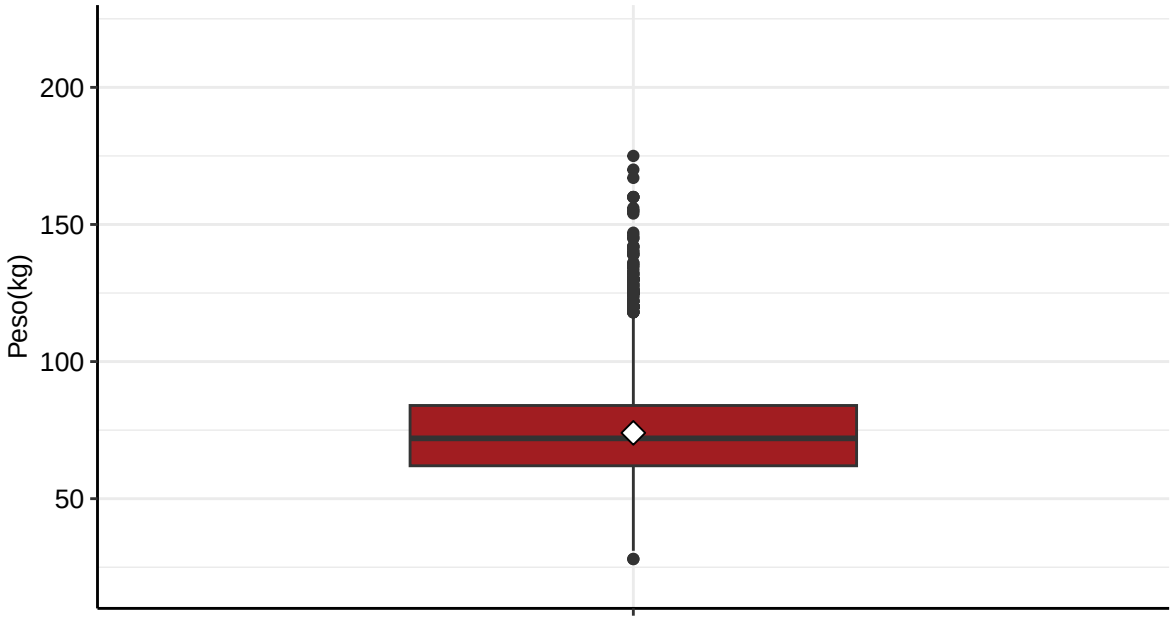

Figura 8: Gráfico boxplot do peso

Quadro 6: Medidas resumo do peso

Estatística	Valor
Média	74,00
Desvio Padrão	16,26
Variância	264,26
Mínimo	28,00
1º Quartil	62,00
Mediana	72,00
3º Quartil	84,00
Máximo	175,00

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

Quadro 7: P-valor do teset de Kolmogorov-Smirnov para Peso

Variável	P-valor	Decisão do teste
Peso	<0,010	Rejeita H_0

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

Quadro 8: P-valor do teste de Kolmogorov-Smirnov para Altura

Variável	P-valor	Decisão do teste
Altura	<0,010	Rejeita H_0

Com a **Figura 6**, podemos observar que há correlação positiva, ou seja quando o peso aumenta, a altura aumenta também. No entanto, à medida que essas variáveis aumentam, os pontos começam a se dispersar, evidenciando que há atletas com baixo peso e maior altura e grande altura e baixo peso, mas a tendência geral dos dados é uma correlação positiva entre peso e altura. O gráfico também mostra que a maioria das pessoas se encontra em uma faixa específica de peso e altura, com uma dispersão menor de dados nessa faixa. Isso sugere que a relação entre peso e altura é relativamente consistente para a maioria dos indivíduos.

Na **Figura 7**, vemos que a altura tem distribuição equilibrada, mas com muitos outliers, percebendo que as medidas dos quartis não variam muito, no entanto há atletas com alturas discrepantes de mínimo e máximo. Isso se vê no quadro de medidas resumo da variável peso, com medidas de variância pequenas, mas valores de mínimo e máximo muito longe da média. Na **Figura 8**, vemos que o peso tem alta variabilidade e muitos outliers além do valor máximo, percebendo que as medidas dos quartis variam muito. Isse se vê no quadro de medidas resumo da variável altura, com medidas de variância altas e valores de mínimo e máximo muito longe da média, sendo a diferença do mínimo para a média de 46 kg e do máximo para média de 101 kg.

Para analisar a relação da altura e peso, foi calculado o coeficiente de correlação de Pearson para peso e altura, com valor igual a 0.790, o que indica uma correlação positiva forte entre altura e peso.

4 Conclusões

Os resultados sugerem que, para otimizar o desempenho dos atletas de elite, que participaram das olimpíadas dos anos de 2000 até 2016, o melhor país para preparar atletas mulheres medalhistas é os Estados Unidos, seguido de Rússia, China, Austrália e Alemanha. Analisando os IMCs dos atletas de ginástica, futebol, judô, atletismo e badminton, vê-se que há diferença entre os IMCs de cada esporte, sendo ginástica o esporte com menor média, tendo 22,58% ginastas abaixo do peso e judô com a maior, observando que nesse esporte 42,5% dos atletas estão acima do peso normal, ou seja o esporte do atleta influencia no seu IMC. Além disso, Michael Fred Phelps, Natalie Anne Coughlin e Ryan Steven Lochte, todos nadadores, são os maiores medalhistas por quantidade de cada tipo de medalha, sendo que há relação entre o medalhista e o tipo de medalha conquistada. Também podemos concluir que o peso e altura dos atletas são variáveis relacionadas, ou seja, quando a altura aumenta, o peso do atleta também aumenta.