Estatística descritiva

Síntese numérica

Prof. Igor Nascimento

12 de agosto de 2021

1 Introdução

Este material apresenta medidas para descrever e sintetizar informações estatísticas, baseado no livro de referência Morettin e Bussab (2017), que é referência para nossa disciplina.

Para compreender as medidas a seguir, considere como **conjunto de dados** os valores ou informações coletadas ou observadas. Cada valor será representado por letras maiúsculas (X, Y ou Z, por exemplo) com índice embaixo indicando a ordem ou a posição. Considere que existam n informações, representadas pela letra X. Dessa forma, o conjunto de informação é $\{X_1, X_2, X_3, ..., X_n\}$.

Vamos considerar que todos os dados são oriundos de uma análise da **população**. Futuramente veremos a diferença em analisar uma **amostra**.

2 Medidas de tendência central

As medidas de tendência central, como o nome sugere, sintetizam os dados informando a centralidade dos dados. São elas: média, mediana e moda.

2.1 Média

2.1.1 Média artimética

A média é a medida mais utilizada pare resumir os dados e pode ser considerada uma medida justa. Uma definição, informal, para média é o valor pelo qual todos os demais podem ser substituídos de maneira que todos sejam iguais.

Considere $n = 6$ dados coletados:

$$X = \{13, 18, 16, 12, 20, 12\}$$

A média dos dados é calculada fazendo:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para os dados de exemplo:

$$\mu = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{13 + 18 + 16 + 12 + 20 + 12}{6} = \frac{91}{6} = 15.16667$$

2.1.2 Média dados agrupados (média ponderada)

Quando os dados são agrupados em k categorias o cálculo é feito por meio da média ponderada pela frequência (f_i repetições dos números):

$$\mu = \frac{\sum_{i=1}^{k} f_i \times x_i}{\sum_{i=1}^{k} f_i} \tag{2.1}$$

sendo $\sum_{i=1}^{k} f_i = n$

Considerando os dados $X = \{13, 18, 16, 12, 20, 12\}$, temos que:

x_i	f_i
12	2
13	1
16	1
18	1
20	1

$$\mu = \frac{12 \times 2 + 13 \times 1 + 16 \times 1 + 18 \times 1 + 20 \times 1}{2 + 1 + 1 + 1} = 15.16667 \tag{2.2}$$

Quando os dados são agrupados em k categorias intervalares:

$$\mu = \frac{\sum_{i=1}^{k} f_i \times \mu_i}{\sum_{i=1}^{k} f_i} \tag{2.3}$$

sendo $\sum_{i=1}^{k} f_i = n$ e μ_i o ponto médio do intervalo.

Por exemplo, considere os seguintes dados em categorias intervalares:

x_i	f_i	μ_i
$[10 - 20)$	2	15
$[20 - 30)$	1	25
$[30 - 40)$	1	35
$[40 - 50)$	1	45
$[50 - 60)$	1	55

Para esses dados, utilizando a média de cada intervalo, temos o seguinte cálculo:

$$\mu = \frac{15 \times 2 + 25 \times 1 + 35 \times 1 + 45 \times 1 + 55 \times 1}{2 + 1 + 1 + 1} = \frac{190}{6} = 31.66667 \tag{2.4}$$

2.1.3 Média geométrica

A média geométrica é utilizada para estudos sobre juros compostos, no contexto de dados financeiros e econômicos. Além disso, também é utilizada para análise de inflação.

Considere $n = 6$ dados:

$$X = \{1.13, 1.18, 1.16, 1.12, 1.20, 1.12\}$$

A média geométrica dos dados é calculada fazendo:

$$\mu_{geo} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

O termo $\prod_{i=1}^{n} x_i$ indica o produtório $X_1 \cdot X_2 \cdot X_3 .... \cdot X_n$.

Para os dados de exemplo:

$$\mu_{geo} = \sqrt[n]{\prod_{i=1}^{n} X_i} = \sqrt[6]{1.13 \cdot 1.18 \cdot 1.16 \cdot 1.12 \cdot 1.20 \cdot 1.12} = \sqrt[6]{2.3282} = 1.1512$$

Lembrete: Para operacionalizar $\sqrt[v]{x}$ na calculadora lembre-se que:

$$\sqrt[a]{x^b} = x^{\frac{b}{a}}$$

Nesse caso, temos:

$$\sqrt[6]{x} = \sqrt[6]{x^1} = x^{\frac{1}{6}}$$

2.2 Mediana (dados desagrupados)

É a medida que indica o valor que está localizado no centro dos dados organizados de forma ordenada. Considere os valores $\{X_1, X_2, X_3, ..., X_n\}$. A representação desses valores de forma ordenada é feita com o uso de colchete no número de índice de cada valor $X_{[i]}$. Isto é, o valor $X_{[i]}$ indica o valor que está na i-ésima posição dos dados **ORDENADOS**.

Além disso, para determinar a mediana em dados desasgrupados é preciso saber se o número de valores n é ímpar ou par.

Caso seja ímpar:

$$\tilde{X} = X_{\left[\frac{n+1}{2}\right]} \tag{2.5}$$

Caso seja par:

$$\tilde{X} = \frac{X_{\left[\frac{n}{2}\right]} + X_{\left[\frac{n}{2}+1\right]}}{2} \tag{2.6}$$

Considere $n = 6$ dados coletados:

$$X = \{13, 18, 16, 12, 20, 12\}$$

Para calcular a mediana desses dados utilizaremos a fórmula para o número par de elementos. Além disso, os dados ordenados são:

$$X = \{12, 12, 13, 16, 18, 20\}$$

$$\tilde{X} = \frac{X_{\left[\frac{6}{2}\right]} + X_{\left[\frac{6}{2}+1\right]}}{2} = \frac{X_{[3]} + X_{[4]}}{2} = \frac{13 + 16}{2} = 14.5 \tag{2.7}$$

2.3 Moda

É o número com maior frequência no conjunto de dados, isto é, é o número que mais aparece. A moda pode não ser única, caso dois ou mais número tenham frequência igual. No caso de duas, o conjunto de dados é considerado bimodal. Caso todos os números apareçam o mesmo número de vezes, não há moda, sendo considerado amodal.

Considerando os dados $X = \{13, 18, 16, 12, 20, 12\}$ é fácil perceber que a moda dos dados é 12.

2.4 Sensibilidade das medidas

A média, mediana e moda se comportam de maneira distanta à presença de valores extremos, isto é, valores muito grandes ou muito pequenos, que fogem do padrão observado dos dados.

Considere adicionar o valor de 1.000.000 aos valores utilizados anteriormente $X = \{13, 18, 16, 12, 20, 12\}$. Qual é a nova média? Qual a nova mediana? Qual a nova moda?

A nova média é:

$$\mu = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{13 + 18 + 16 + 12 + 20 + 12 + 1.000.000}{7} = \frac{1.000.091}{7} = 142.870,1$$

Dessa forma, apenas 1 valor (extremo) alterou a média de 15.1 para 142.870,1. Por isso, dizemos que a **média é muito sensível à valores extremos**.

A nova mediana é calculada utilizando a fórmula do número ímpar de valores:

$$\tilde{X} = X_{\left[\frac{n+1}{2}\right]} = X_{\left[\frac{7+1}{2}\right]} = X_{[4]} = 16$$

Dessa forma, apenas 1 valor (extremo) alterou a mediana de 14.5 para 16. Por isso, dizemos que a **mediana é muito pouco sensível à valores extremos**.

Percebe-se ainda, que, apesar do valor extremo adicionado, o valor da moda não foi alterado. Dessa forma, dizemos que a moda não é afetada por valores extremos.

3 Medidas de dispersão

3.1 Amplitude

A amplitude do conjunto de dados indica qual o intervalo total de variabilidade dos dados, isto é, o intervalo entre o valor mínimo ($X_{[1]}$) e o valor máximo ($X_{[n]}$).

$$A = X_{[n]} - X[1]$$

Considerando os dados $X = \{13, 18, 16, 12, 20, 12\}$ a amplitude dos dados é $20 - 12 = 8$. Dessa forma, os dados tem uma amplitude de 8, isto é, os dados variam no intervalo de 8 unidades.

Essa medida é **extremamente afetada por valores extremos**. Caso adicionemos o valor de 1.000.000 a amplitude passa a ser $1.000.000 - 12 = 9.999.988$. Nesse caso e na grande maioria dos casos, a amplitude não reflete corretamente a variabilidade dos dados.

3.2 Variância

Diferentemente da amplitude, a variância (σ^2) considera uma amplitude média, que compara cada valor com um valor de referência, que é a média dos dados. Dessa forma cada valor é comparado com a média, e a diferença entre essa comparação é utilizada para determinar a variabilidade média.

A variância dos dados (populacionais) é calculada fazendo:

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

No entanto, com algum esforço matemático é possível obter uma **fórmula simplificada da variância**, que utiliza apenas somatórios.

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

Para os dados de exemplo:

$$\sigma^2 = \frac{\sum_{i=1}^{6} x_i^2}{6} - \left(\frac{\sum_{i=1}^{6} x_i}{6}\right)^2 = \frac{13^2 + 18^2 + 16^2 + 12^2 + 20^2 + 12^2}{6} - \left(\frac{91}{6}\right)^2 = \frac{1437}{6} - \frac{8281}{36} = 9.47222$$

3.3 Desvio padrão

O desvio padrão é a raiz da variância e é mais adequado do que a variânça, pois está na escala dos dados:

$$\sigma = \sqrt{\sigma^2} = \sqrt{9.4722} = 3.07$$

3.4 Coeficiente de variação

O coeficiente de variação (CV) é a medida correta para comparar variabilidade de dados diferentes. Essa medida é calculada utilizando a média μ e o desvio-padrão σ dos dados.

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Dessa forma, o CV indica o percentual de variabilidade em relação a média dos dados.

Para compreender essa medida, considere as seguintes situação:

A Gasto com cartão de crédito: O gasto média com cartão de crédito de uma certa família é de R$ 2.000,00 e o desvio padrão é de R$ 100,00.

B Gasto com energia elétrica: o gasto médio com energia elétrica em uma certa residência é de R$ 300,00 e o desvio padrão é de R$ 75,00

Qual das duas despesas, cartão de crédito ou energia elétrica, **há maior variabilidade**?

Se analizarmos apenas o desvio padrão σ, o gasto com cartão é maior, pois de R$ 100,00 é maior do que R$ 75,00.

No entanto, para respondermos de forma correta precisamos considerar o padrão médio dos dados, utilizando o coeficiente de variação (CV).

No caso de A, a variabilidade é de R$ 100 em torno R$ 2.000,00. No caso de B, a variabilidade é de R$ 75,00 em torno de R$ 300,00. Os coeficientes de variação são:

$$CV_A = \frac{\sigma_A}{\mu_A} \cdot 100\% = \frac{100}{2000} \cdot 100\% = 5\%$$

$$CV_B = \frac{\sigma_B}{\mu_B} \cdot 100\% = \frac{75}{300} \cdot 100\% = 25\%$$

Assim, o gasto com cartão de crédito possui uma variabilidade de 5% do valor médio gasto. Enquanto a despesa com energia elétrica tem variabilidade de 25% do valor médio observado.

Dessa forma, o gasto com energia elétrica, baseado no coeficiente de variação, possui maior variabilidade do que o gasto com cartão de crédito.

4 Valor Padrão

O valor padrão (ou valor Z) serve para comparar valores de diferentes padrões de dados. Para isso, utiliza-se a média μ e o desvio padrão σ do conjunto de dados de onde foi retirado a informação. O valor padrão do valor x, que pertence a um conjunto de dados com média μ e desvio padrão σ é calculado por:

$$Z = \frac{x - \mu}{\sigma}$$

O numerador indica se o valor x está acima ou abaixo da média. Se o valor for positivo, está acima, se for negativo, está abaixo. O denominador indica o quão distante está x da média, em unidades de desvio padrão. Dessa forma, o valor Z permite comparar valor de diferentes conjuntos de dados.

Para compreender essa medida, considere as seguintes situação:

A O aluno A tirou nota $x = 5$ em uma turma em que a nota média foi $\mu_A = 4$ e com desvio padrão de $\sigma_A = 1$.

B O aluno B tirou nota $x = 6$ em uma turma em que a nota média foi $\mu_B = 7$ e com desvio padrão de $\sigma_B = 2$.

Qual dos dois alunos, teve **maior nota**?

Para responder corretamente podemos utilizar o valor Z para a nota do aluno A (Z_A) e do aluno B (Z_B).

$$Z_A = \frac{A - \mu_A}{\sigma_A} = \frac{5 - 4}{1} = \frac{+1}{+1} = +1$$

$$Z_B = \frac{B - \mu_B}{\sigma_B} = \frac{6 - 7}{2} = \frac{-1}{+2} = -0.5$$

Comparando apenas as notas, temos que a nota do aluno B de 6 é maior do que a nota do aluno A que foi de 5. No entanto, considerando a média, temos que o aluno A teve nota ACIMA DA MÉDIA da turma em 1 ponto. Já o aluno B, a nota está ABAIXO DA MÉDIA da turma em 1 ponto. Isso pode ser visto analisando o numerador do valor Z.

Além disso, precisamos comparar essa distância com o desvio padrão. Para o caso do aluno A, as notas a turma têm desvio padrão de 1 ponto. Nesse caso, pode-se dizer que o aluno A está **1 desvio padrão acima da média**.

Para o caso do aluno B, as notas a turma têm desvio padrão de 2 pontos. Nesse caso, pode-se dizer que o aluno B está **0.5 desvio padrão abaixo da média**.

Por fim, temos que que $Z_A > Z_B$, isto é, o valor Z para a nota do aluno A é maior do que o valor Z do aluno B, então a nota do aluno A foi **MAIOR DO QUE** a do aluno B.

Dessa forma, de agora em diante, quando for comparar valores de conjunto de dados diferentes, é preciso considerar o padrão dos dados, que envolve analisar a **média e o desvio padrão** dos dados. Sempre que for confrontado com a pergunta "o que é maior, X ou Y?"Responda: "Depende!! Me informe a média e o desvio padrão dos dados"

5 Quartis

5.1 Dados desagrupados

O cálculo de percentis para dados desagrupados é um pouco mais simples quando comparada com os dados desagrupados, mas não é única. É importante lembrar que os dados precisam estar ordenados (rol). Vamos considerar a seguinte fórmula para o cálculo do percentil $p\%$:

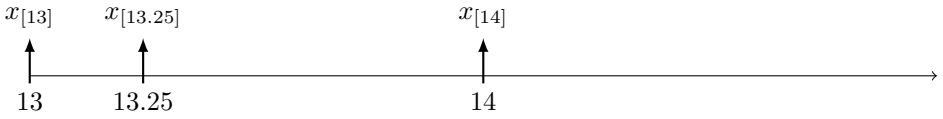
$$k = p\% \times (n + 1)$$

sendo $p\%$ o percentil que queremos, n o número de elementos e k a posição do elemento que queremos $x_{[k]}$. Lembre-se que $p = 25\%$ é o primeiro quartil (Q1), assim como $p = 75\%$ é o terceiro quartil (Q3).

Considere o caso de $p = 25\%$ e $n = 52$. Então, temos que:

$$k = 0.25 \times (52 + 1) = 13.25$$

Nesse caso, queremos o valor $x_{[13.25]}$. Perceba que $x_{[13.25]}$ está entre o $x_{[13]}$ e $x_{[14]}$, porém mais próximo de $x_{[13]}$. Nesse caso, assim como é feito com a mediana, não podemos tirar a média, pois não seria justo com o $x_{[13]}$.



Então faremos uma média ponderada baseado na diferença entre os intervalos acima:

$$\frac{x_{[13.25]} - x_{[13]}}{13.25 - 13} = \frac{x_{[14]} - x_{[13]}}{14 - 13} \tag{5.1}$$

Resolvendo a relação acima, temos que:

$$x_{[13.25]} = x_{[13]} + 0.25 \times (x_{[14]} - x_{[13]})$$

Dessa forma, $x_{[13]}$ tem um peso maior por estar mais próximo de $x_{[13.25]}$.

$$x_{[13.25]} = x_{[13]} \times 0.75 + 0.25 x_{[14]}$$

5.2 Dados agrupados

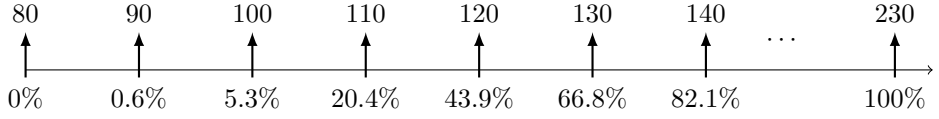
Os dados a seguir apresentam a ditribuição de frequência em intervalo de classes.

Faixa	FA ou Fi	FR	FRA
[80-90)	52		
[90-100)	420		
[100-110)	1335		
[110-120)	2083		
[120-130)	2028		
[130-140)	1354		
[140-150)	748		
[160-170)	240		
[150-160)	318		
[180-190)	72		
[170-180)	138		
[210-220)	13		
[200-210)	21		
[190-200)	28		
[220-230)	5		

O cálculo de quartis nesse caso é um pouco diferentes, e utiliza a frequência relativa acumulada (FRA). Como o nome sugere, a FRA é obtida acumulando os percentuais da frequência relativa (FR) a medida que aumenta o intervalo de classe. Vale lembrar que a frequência relativa da primeira classe, $[80 - 90)$, é obtida dividindo a frequência absoluta (FA) pelo total $52/8855 = 0.6\%$.

Faixa	FA ou Fi	FR	FRA
[80-90)	52	0,6%	0,6%
[90-100)	420	4,7%	5,3%
[100-110)	1335	15,1%	20,4%
[110-120)	2083	23,5%	43,9%
[120-130)	2028	22,9%	66,8%
[130-140)	1354	15,3%	82,1%
[140-150)	748	8,4%	90,6%
[160-170)	240	2,7%	93,3%
[150-160)	318	3,6%	96,9%
[180-190)	72	0,8%	97,7%
[170-180)	138	1,6%	99,2%
[210-220)	13	0,1%	99,4%
[200-210)	21	0,2%	99,6%
[190-200)	28	0,3%	99,9%
[220-230)	5	0,1%	100,0%

Perceba que a FRA para na segunda classe, $[90 - 100)$, é obtida somando as FR da primeira e segunda classe $0.6\% + 4.7\% = 5.3\%$. Dessa forma, temos a seguinte representação para os limites entre as classes e as frequências relativas acumuladas:

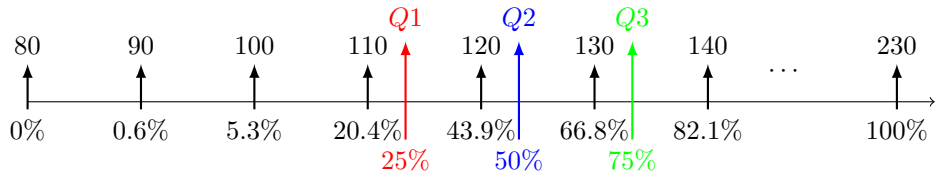


Com base na representação acima é possível percebe que 43.9% dos valores estão abaixo de 120. De maneira análoga, apenas 17.9% dos valores estão acima de 140.

Essas são medidas separatrizes, que indicam as posições dos valores baseado em padrão em que os dados estão distribuídos.

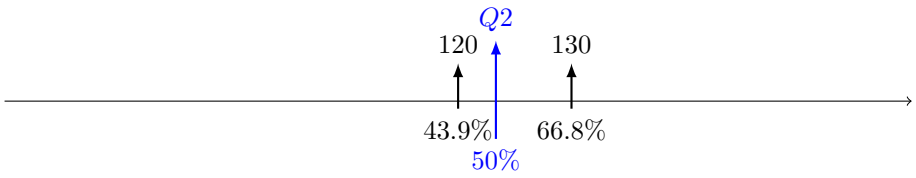
As medidas mais conhecidas são Q1 = 25% (primeiro quartil), Q2 = 50% (segundo quartil ou mediana) e Q3 = 75% (primeiro quartil).

Para encontrar a mediana nesse conjunto de dados, inicialmente é preciso encontrar o intervalo que contém a mediana. Para isso, é preciso utilizar a FRA. Perceba que a mediana é o valor que está na posição 50% no conjunto de dados.

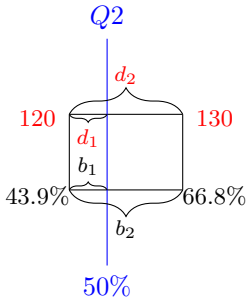


Faixa	FA ou Fi	FR	FRA
[80-90)	52	0,6%	0,6%
[90-100)	420	4,7%	5,3%
[100-110)	1335	15,1%	20,4%
[110-120)	2083	23,5%	43,9%
[120-130)	2028	22,9%	66,8%
[130-140)	1354	15,3%	82,1%
[140-150)	748	8,4%	90,6%
[160-170)	240	2,7%	93,3%
[150-160)	318	3,6%	96,9%
[180-190)	72	0,8%	97,7%
[170-180)	138	1,6%	99,2%
[210-220)	13	0,1%	99,4%
[200-210)	21	0,2%	99,6%
[190-200)	28	0,3%	99,9%
[220-230)	5	0,1%	100,0%

Com isso, o cálculo do segundo quartil (mediana) pode ser feito de maneira análoga aos dados desagrupados. Nesse caso, fica evidente que o valor de Q2 está mais próximo a 120 do que de 130, pois 50% está mais próximo de 43.9% do que de 66.8%.



O valor de Q2 pode ser encontrado com maior facilidade por meio da proporcionalidade das distâncias b_1, b_2, d_1 e d_2 apresentado a seguir:



Dessa forma temos:

$$\frac{d_1}{b_1} = \frac{d_2}{b_2} \tag{5.2}$$

$$\frac{Q_2 - 120}{50\% - 43.9\%} = \frac{130-120}{66.8\%-43.9\%} \tag{5.3}$$

$$\frac{Q_2 - 120}{6.1\%} = \frac{10}{22.9\%} \tag{5.4}$$

$$(Q_2 - 120) \times 22.9\% = 10 \times 6.1\% \tag{5.5}$$

$$Q_2 \times 22.9\% - 120 \times 22.9\% = 0.61 \tag{5.6}$$

$$Q_2 \times 22.9\% - 27.5 = 0.61 \tag{5.7}$$

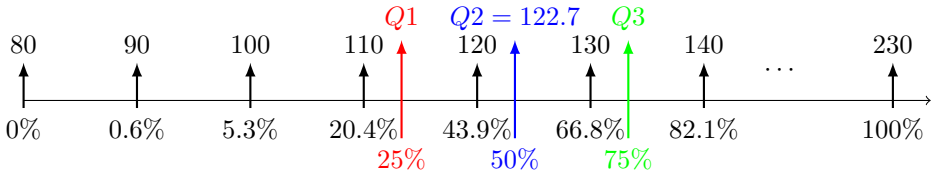
$$Q_2 \times 22.9\% = 28.11 \tag{5.8}$$

$$Q_2 = \frac{28.11}{0.229} \tag{5.9}$$

$$Q_2 = 122.7 \tag{5.10}$$

$$\tag{5.11}$$

Dessa forma, 122.7 é o valor que divide o conjunto de dados em metades iguais. É possível dizer, então, que 50% dos valores estão abaixo de 122.7 ou que metade dos valores estão acima de 122.7.



6 Histograma

Como vimos anterioremente, os dados podem ser apresentados de forma desagregada, valor por valor, ou por meio de intervalos. No caso da distribuição agregada em intervalos há uma nova informação: a **frequência absoluta** (FA). A variável se transforma em um intervalo, e o valor da frequência desse intervalo é feito contabilizando o número de valores desagregados que se enquadram em cada intervalo. A tabela 1 a seguir mostra um exemplo de criação de intervalos de frequência.

Faixa	FA ou Fi	FR	FRA	Densidade
[80-90)	52	0,6%	0,6%	0,001
[90-100)	420	4,7%	5,3%	0,005
[100-110)	1335	15,1%	20,4%	0,015
[110-120)	2083	23,5%	43,9%	0,024
[120-130)	2028	22,9%	66,8%	0,023
[130-140)	1354	15,3%	82,1%	0,015
[140-150)	748	8,4%	90,6%	0,008
[160-170)	240	2,7%	93,3%	0,003
[150-160)	318	3,6%	96,9%	0,004
[180-190)	72	0,8%	97,7%	0,001
[170-180)	138	1,6%	99,2%	0,002
[210-220)	13	0,1%	99,4%	0,000
[200-210)	21	0,2%	99,6%	0,000
[190-200)	28	0,3%	99,9%	0,000
[220-230)	5	0,1%	100,0%	0,000
Total	8855	100,0%	-	-

Figura 1 – Dados agregados em intervalos

O número de barras ou intervalos (k) pode ser determinado de diversas maneiras, no entanto a mais habitual e apresentada em Morettin e Bussab (2017) é por meio da fórmula de Sturges, baseada no número de observações (n):

$$k = 3.322 \times log(n) + 1 \tag{6.1}$$

Além da FA também pode-se utilizar a frequência relativa (FR) para representar os dados. Como exemplo, a frequência relativa da primeira classe, $[80 - 90)$, é obtida dividindo a frequência absoluta (FA) pelo total $52/8855 = 0.6\%$. Como dito anteriormente, a frequência relativa acumulada (FRA) é obtida acumulando os percentuais da FR a medida que aumenta o intervalo de classe. Outra medida associada a represenção em intervalos é a de densidade, que divide o valor da frequência relativa pelo tamanho do intervalo das faixas. Por exemplo, a densidade da primeira categoria é $52/8855/10 = 0.001$.

O histograma é obtido pela representação gráfica da desses intervalos por meio do gráfico de coluna, sendo a altura das barras iguais à frequência absoluta, frequência relativa ou densidade.

Uma boa prática é adotar intervalos igualmente espaçados para facilitar a comparação entre as frequências. Além disso, na prática, os intervalos de apresentação dos dados são normalmente com números "fáceis" múltiplos de 2, 5 ou 10.

O histograma facilita a análise da **distribuição** dos dados, isto é, como os dados se distribuiem ao longo dos valores possíveis.

7 Boxplot

O boxplot (gráfico de caixa) é um gráfico utilizado para representar a distribuição dos dados baseado nas medidas

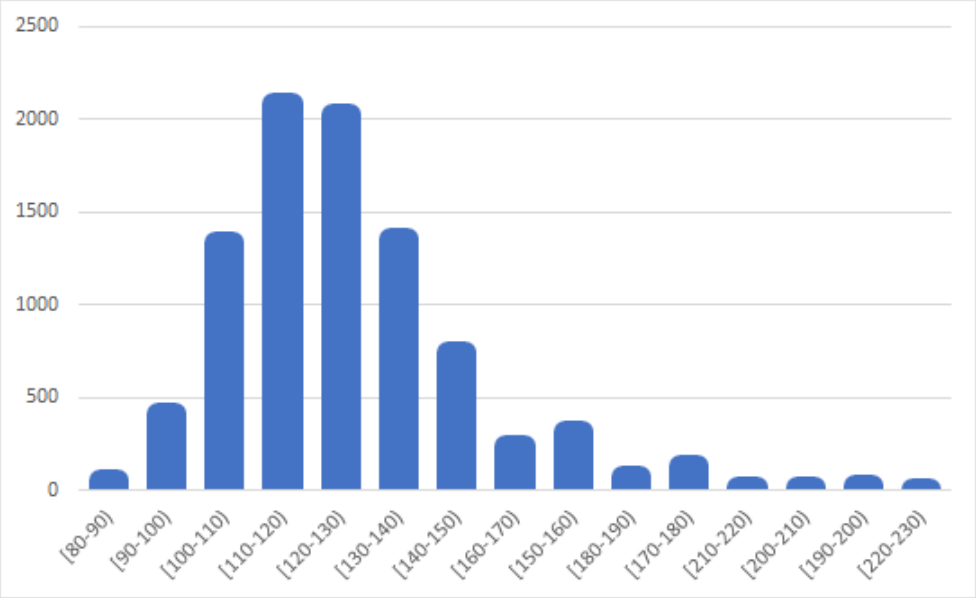


Figura 2 – Histograma com a frequência absoluta

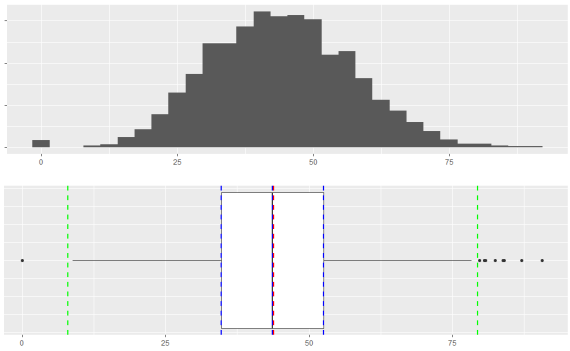


Figura 3 – Histograma e boxplot

dos quartis.

A figura 3 apresenta duas representações de um conjunto de dados, beseado no histograma e no boxplot.

Cada posição da caixa está relacionada a um quartil dos dados. A figura 4 apresenta a indicação de cada medida de posição, sendo Q1 o primeiro quartil, Q2 o segundo quartil (mediana), Q3 o terceiro quartil, LS o limite superior e LI o limite inferior.

$$LS = Q3 + 1.5 \times (Q3 - Q1)$$

$$LI = Q1 - 1.5 \times (Q3 - Q1)$$

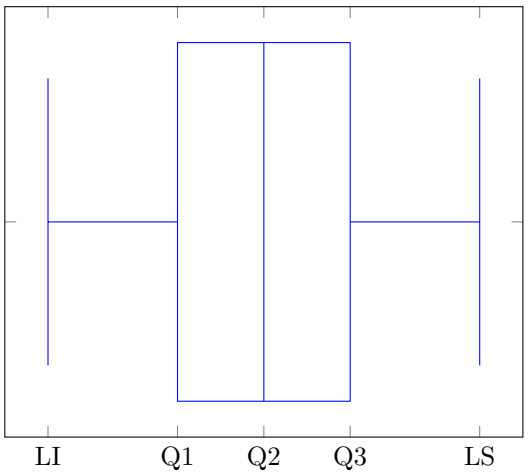


Figura 4 – Configuração do Boxplot

8 Assimetria e curtose

Medida de assimetria fornece informações sobre o formato da distribuição dos dados. Para compreeender o que é assimetria, inicialmente vamos compreender o que simetria.

Segundo o dicionário, simetria é uma correspondência em tamanho, forma ou arranjo, de partes dispostas em lados contrários de uma linha divisória, um plano, um centro ou um eixo. Dessa forma, precisamos de uma referência para determinar a simetria.

Analisando a imagem 2 perceba que a distribuição ou frequência do extremo do lado direito não é igual ao do lado esquerdo, independentemente do ponto escolhido para análise. Um exemplo disso, é escolhendo a classe modal de 110 a 120. Perceba que o lado direito da imagem há 11 intervalos, enquanto que o esquerdo tem apenas 3. Dessa forma, essa imagem pode ser considerada assimétria, isto é, a ausência de simetria.

Na estatística, assim como mensuramos o valor médio e a variabilidade dos dados, podemos medir o nível de simetria dos dados. Essa medida pode ser feita de três maneiras, e para todas elas utilizamos medidas de tendência central e medidas de variabilidade.

A primeira medida (AS_1) utiliza a diferênça entre a média (μ) e a moda (M_0) dividada pelo desvio padrão (σ). Quanto maior a distância entre a média e a moda, menor será a simetria e maior será a assimetria.

$$AS_1 = \frac{\mu - M_0}{\sigma} \tag{8.1}$$

A segunda medida (AS_2) utiliza a diferênça entre a média (μ) e a mediana (Q_2) dividada pelo desvio padrão (σ). Quanto maior a distância entre a média e a moda, menor será a simetria e maior será a assimetria.

$$AS_2 = \frac{3(\mu - Q2)}{\sigma} \tag{8.2}$$

Para compreender as medidas AS_1 e AS_2 veja a imagem 5. Essa distribuição de dados apresenta um comportamento simétrico e nesse tipo de situação, o valor da média, moda e mediana coincidêm, isto é, todas as medidas de tedência central estão no mesmo ponto.

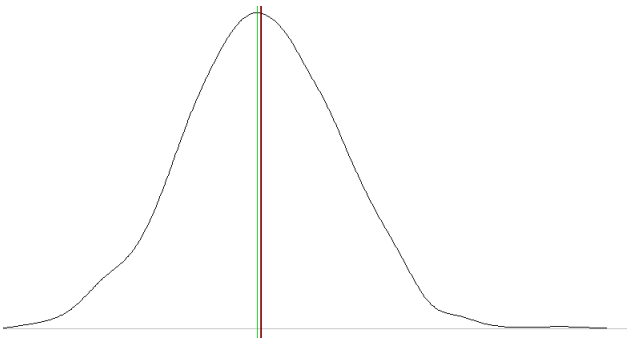


Figura 5 – Distribuição simétrica $AS_2 = 0$, Moda=Mediana=Média

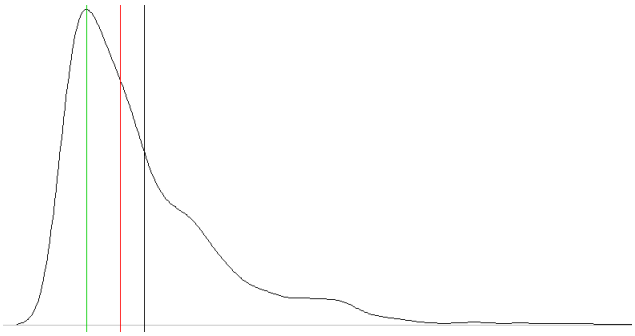


Figura 6 – Distribuição com assimetria positiva, Moda¡Mediana¡Média

Existem dois tipos de ausência de simetria. A primeira delas é uma assimetria postiva ou assimetria à direita.

Veja a imagem 6 e perceba que há muitos valores altos do lado direito da distribuição, isto é, há frequência observada no lado direito. Esses valores podem ser considerados valores extremos à medida que se distanciam muito da região onde os dados estão concentrados. Observe que cada uma das linhas coloridas representa uma medida de tendência central. Pela definição, fica fácil perceber que a linha verde representa a moda, pois é o valor/ponto/posição (eixo x) com maior frequência (eixo y), isto é, ponto mais alto do histograma. Como vimos no caso da média, ela é muito afetada por valores extremos, podemos inferir que a linha preta representa a média. Isso acontece devido os valores do lado direito **elevarem** o valor da média, isto é, "puxarem" a média para cima. Por fim e por eliminação, a linha vermelha é a mediana, que divide extamente os dados em duas metades iguais, isto é, 50% dos valores ou observações estão antes da linha vermelha e metade após a linha vermelha.

Nesse tipo de assimetria à direita ou positiva, perceba que a média é maior do que a mediana que é maior do que a moda. Dessa forma, analisando a equação 8.1, percebemos que o numerador será positivo, pois a média é maior do que a moda. De maneira análoga, analisando a equação 8.2, percebemos que o numerador será positivo, pois a média é maior do que a mediana. Portanto, em qualqer dessas situação o valor de assimetria será **positivo**. Para fixar isso, lembre-se **ASSIMETRIA À DIREITA TEM MAIS VALORES MAIORES, PORTANTO TERÁ ASSIMETRIA POSITIVA**.

Veja a imagem 7 e perceba que há muitos valores altos do lado esquerdo da distribuição, isto é, há frequência observada no lado esquerdo. Esses valores podem ser considerados valores extremos à medida que se distanciam muito da região onde os dados estão concentrados. Observe que cada uma das linhas coloridas representa uma medida de tendência central. Pela definição, fica fácil perceber que a linha verde representa a moda, pois é o valor/ponto/posição (eixo x) com maior frequência (eixo y), isto é, ponto mais alto do histograma. Como vimos

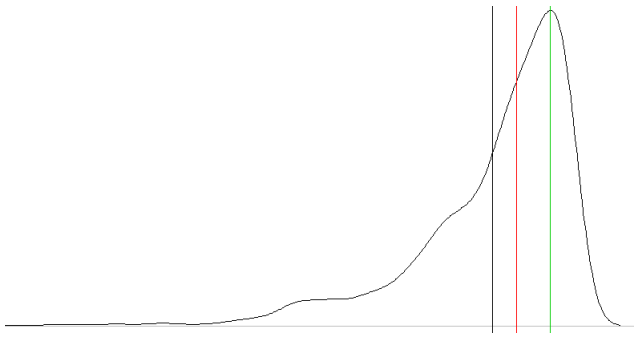


Figura 7 – Distribuição com assimetria negativa, Moda¿Mediana¿Média

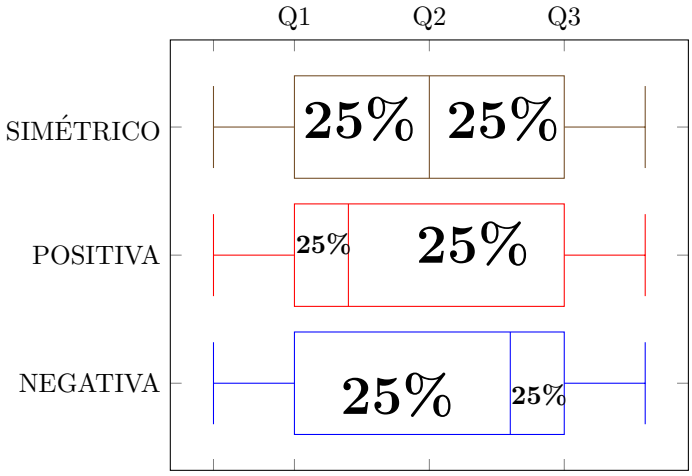


Figura 8 – Distribuição dos dados segundo tempo de cura.

no caso da média, ela é muito afetada por valores extremos, podemos inferir que a linha preta representa a média. Isso acontece devido os valores do lado direito **elevarem** o valor da média, isto é, "puxarem" a média para baixo. Por fim e por eliminação, a linha vermelha é a mediana, que divide extamente os dados em duas metades iguais, isto é, 50% dos valores ou observações estão antes da linha vermelha e metade após a linha vermelha.

Nesse tipo de assimetria à esquerda ou negativa, perceba que a média é menor do que a mediana que é menor do que a moda. Dessa forma, analisando a equação 8.1, percebemos que o numerador será negativo, pois a média é menor do que a moda. De maneira análoga, analisando a equação 8.2, percebemos que o numerador será negativo, pois a média é menor do que a mediana. Portanto, em qualqer dessas situação o valor de assimetria será **negativa**. Para fixar isso, lembre-se **ASSIMETRIA À ESQUERDA TEM MAIS VALORES MENORES, PORTANTO TERÁ ASSIMETRIA NEGATIVA**.

A terceira e última medida (AS_3) utiliza apenas informações dos quartis Q_1, Q_2 e Q_3.

$$AS_3 = \frac{(Q3 - Q2) - (Q2 - Q1)}{Q3 - Q1} = \frac{Q3 + Q1 - 2 \times Q2}{Q3 - Q1} \tag{8.3}$$

Para compreender a equação 8.3 vamos analisar os seguintes boxplots da figura 8. Analisando o boxplot com legenda de **simétrico**, a diferença entre os quartis Q2 e Q1 é a mesma de Q3 para Q2. Dessa forma, o numerador de da equação 8.3 é igual a zero, o que indica que 50% dos dados são simétricos em torno de Q2.

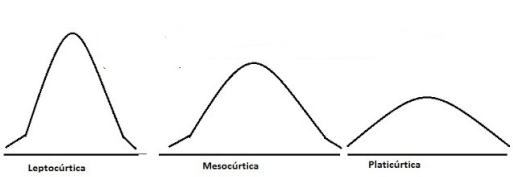


Figura 9 – Classificações quanto a curtose.Fonte:google

Analisando o boxplot com legenda de **postiva** da figura 8, a diferença entre os quartis Q2 e Q1 é **menor** de Q3 para Q2. Dessa forma, o numerador de da equação 8.3 é **maior** do que zero, o que indica uma assimetria postiva.

De maneira análoga, analisando o boxplot com legenda de **negativa** da figura 8, a diferença entre os quartis Q2 e Q1 é **maior** de Q3 para Q2. Dessa forma, o numerador de da equação 8.3 é **menor** do que zero, o que indica uma assimetria negativa.

Além da assimetria, outra medida relacionado ao formato da distribuição dos dados é a mensuração do achatamento por meio da **curtose**.

$$C = \frac{(Q3 - Q1)/2}{D9 - D1} \tag{8.4}$$

Sendo D9 e Q1 o nono (90%) e o primeiro decil(10%). O valor calculado para a curtose pode ser classificado em e a representação pode ser vista na imagem 9 :

- valores próximos à 0.26367 mesocúrtica
- valores menores do que 0.26367 platocúrtica
- valores maiores do que 0.26367 leptocúrtica

Referências

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. São Paulo: Editora Saraiva, 2017.