

Mini curso de Probabilidade e Estatística

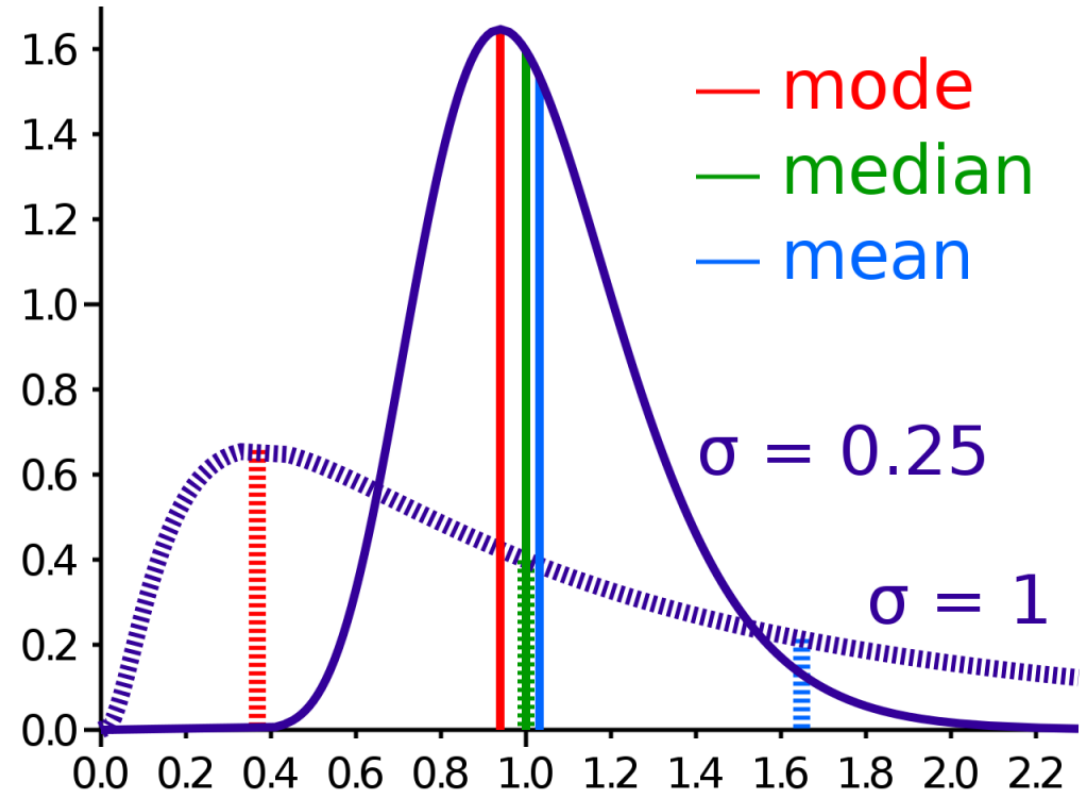
Por tutor Mestre Omar Barroso Khodr

Aula 2

- Média, mediana e moda
- Variância e desvio padrão
- Assimetria e curtose
- Representação gráfica dos dados

Centros de distribuição

- O centro de uma distribuição é geralmente chamado de valor mais típico da distribuição. Existem três maneiras de descrever o centro de uma distribuição: **mediana, média e moda.**



Mode = Moda; Median = Mediana; Mean = Média. Fonte: Albert.io

Mediana

- Aonde é o centro de uma régua? O centro pode ser visto como o ponto de equilíbrio que corta a régua em duas metades com pesos iguais. Por exemplo, o centro de uma régua de 30cms é 15cms.
- Uma ideia semelhante pode ser aplicada ao centro de uma distribuição: o centro de uma distribuição pode ser visto como o valor que divide os dados ordenados em duas metades com um número igual de observações. Esse valor é conhecido como mediana.

Mediana

- Ou seja, 50% das observações estão abaixo da mediana e outros 50% estão acima dela. Aqui estão os passos para encontrar a mediana de um conjunto de dados:
- 1. Classifique os dados do **menor para o maior**.
- 2. Se o número total de observações n for **ímpar**, a mediana será a observação que está no **meio da lista** ordenada.
- 3. Se n for um número **par**, a mediana será a **média dos dois valores no meio** da lista ordenada.

Mediana

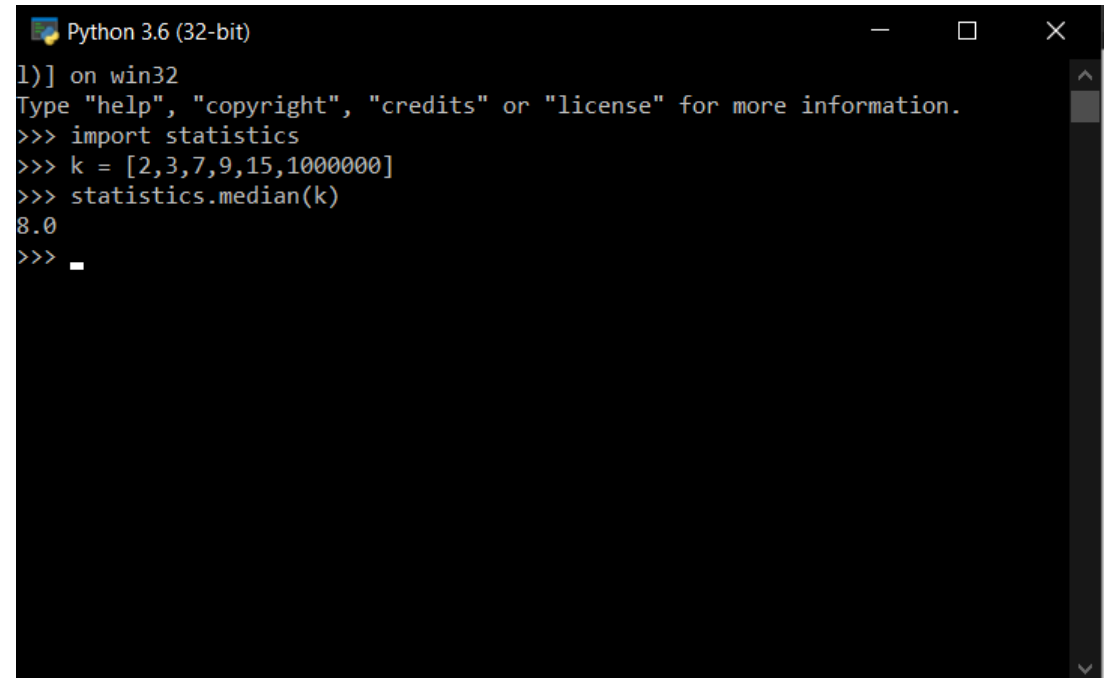
- Por exemplo, temos o nosso **conjunto** 'A' com uma coleção de elementos de determinada maneira:
- $A = (3, 5, 3, 7, 7, 8, 9)$
- Vamos calcular a mediana...
- Primeiro, devemos ordenar os dados de forma ascendente (do menor para o maior)... Desta maneira, nosso **conjunto** ordenado é:
- $C = (3, 3, 5, 7, 7, 8, 9)$
- Agora, podemos facilmente encontrar a **mediana**: 3, 3, 5, **7**, 7, 8, 9.
- Nesse caso, como a coleção dos nossos elementos é composta de forma ímpar, a mediana é encontrada de forma simples no meio do conjunto.

Mediana

- Por exemplo, temos o nosso **conjunto** 'B' com uma coleção de elementos de determinada maneira:
- $B = (2, 3, 7, 15, 9, 1.000.000)$
- Vamos calcular a mediana...
- Primeiro, devemos ordenar os dados de forma ascendente (do menor para o maior)... Desta maneira, nosso **conjunto** ordenado é:
- $K = (2, 3, 7, 9, 15, 1.000.000)$
- Agora, podemos encontrar a **mediana** ao escolher os dois números ao meio do conjunto e calculando a sua média. Ou seja:
- $K = (2, 3, 7, 9, 15, 1.000.000); (7+9)/2 = 16/2 = 8.$

Calculando pelo Python

- 1. Abra o prompt do Python
- 2. Escreva: “import statistics”
- 2.1. Essa é uma biblioteca do Python que importa cálculos de estatística.
- 3. Crie um conjunto: $k = [2,3,7,9,15,1000000]$
- 4. Escreva: “statistics.median(k)”
- 5. Voilá, a mediana é **8**.



```
Python 3.6 (32-bit)
1)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import statistics
>>> k = [2,3,7,9,15,1000000]
>>> statistics.median(k)
8.0
>>> _
```


Média

- Existem vários tipos de médias em estatística, cada uma com seu método de cálculo e finalidade específicos. Aqui estão as mais comuns:

- 1. Média Aritmética (Média)

- **Esta é a média mais conhecida.**

- Calculada somando todos os valores em um conjunto de dados e dividindo pelo número de valores.

- Fórmula:

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- No qual, x_i representa os elementos de um conjunto; $\sum_{i=1}^n (x_1 + x_2 + \cdots + x_n)$ representa a soma dos elementos dentro do conjunto; e por final 'n' representa o número dos elementos que compõe o conjunto (no qual dividimos) pelo somatório.

Média

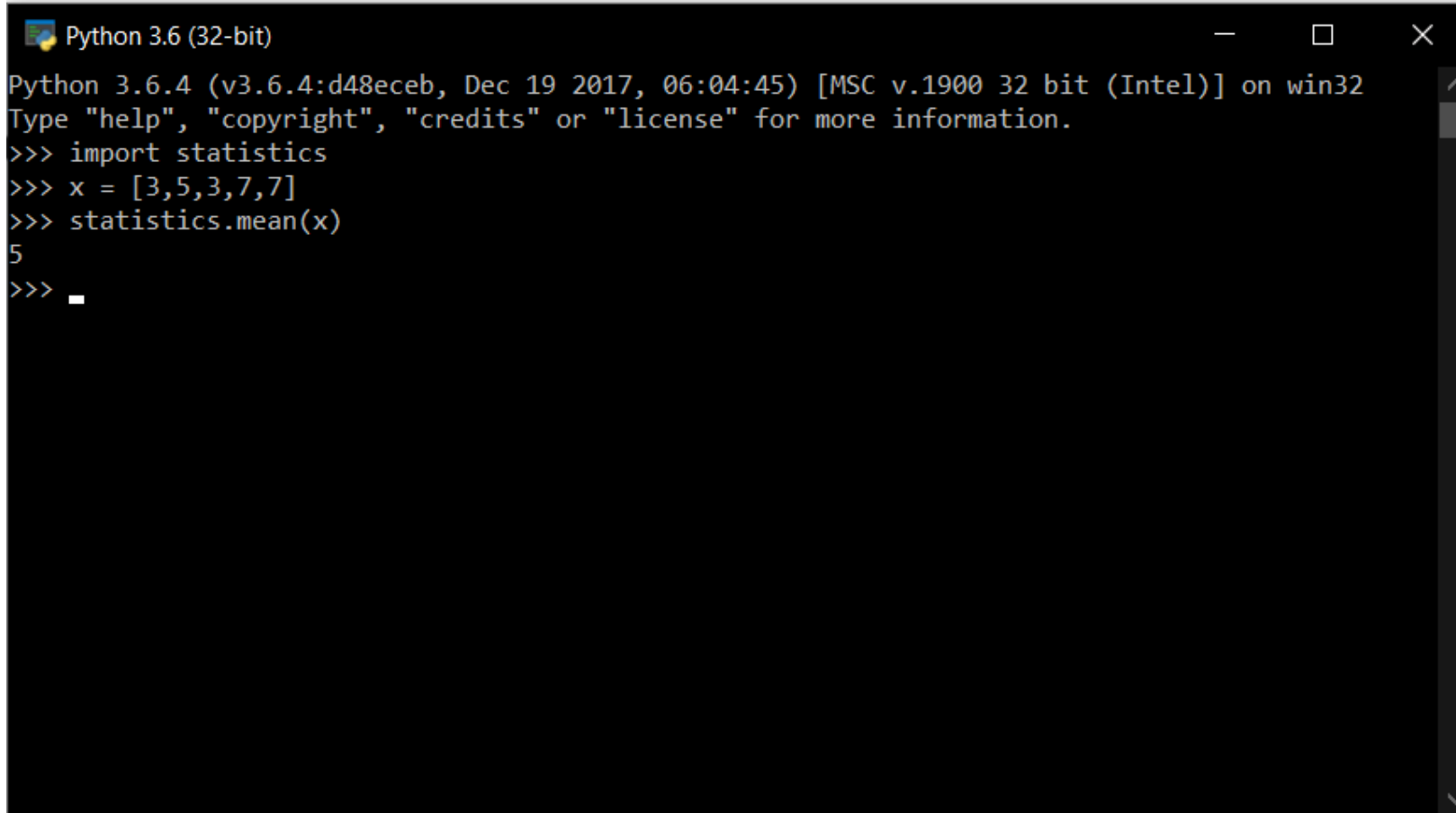
- Exemplo, média aritmética...

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- $x = 3, 5, 3, 7, 7$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+3+7+7}{5} = \frac{25}{5} = 5.$$

Média (python)

A screenshot of a Python 3.6 (32-bit) command prompt window. The window title is "Python 3.6 (32-bit)". The prompt shows the Python version and build information: "Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32". It then shows the user entering commands to import the statistics module, create a list x = [3, 5, 3, 7, 7], and calculate the mean of x, which returns 5. The prompt ends with a cursor on a new line.

```
Python 3.6 (32-bit)
Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import statistics
>>> x = [3,5,3,7,7]
>>> statistics.mean(x)
5
>>> _
```

Média

- 2. Média Geométrica
- Útil ao comparar razões ou porcentagens.
- Calculada multiplicando todos os valores em um conjunto de dados e extraíndo a raiz quadrada de n (onde n é o número de valores).

$$\bar{x} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = (x_1 x_2 \cdots x_n)^{\frac{1}{n}}$$

- Exemplo: Comumente usada em finanças, crescimento populacional ou taxas de crescimento.

$$(4 \times 36 \times 45 \times 50 \times 75)^{\frac{1}{5}} = \sqrt[5]{24\,300\,000} = 30$$

Média

- 3. Média Harmônica
- Usada para conjuntos de dados com taxas ou razões, como velocidade ou eficiência.
- Enfatiza valores menores com mais intensidade do que a média aritmética.

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

- Exemplo: Velocidade média quando as distâncias são constantes.
- Nosso conjunto é composto por: 4, 36, 45, 50, 75

$$\frac{5}{\frac{1}{4} + \frac{1}{36} + \frac{1}{45} + \frac{1}{50} + \frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15$$

Média

- 3. Média Harmônica
- Exemplo: Velocidade média quando as distâncias são constantes.
- Nosso conjunto é composto por: 4, 36, 45, 50, 75

$$\frac{5}{\frac{1}{4} + \frac{1}{36} + \frac{1}{45} + \frac{1}{50} + \frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15$$

- Por exemplo, se tivermos cinco mangueiras que podem esvaziar um tanque de um determinado tamanho em 4, 36, 45, 50 e 75 minutos, então a média harmônica de 15 nos diz que essas cinco mangueiras diferentes trabalhando juntas produzirão na mesma taxa que cinco mangueiras que podem esvaziar o tanque em 15 minutos cada.

Média

- 4. Média Ponderada
- Útil quando valores diferentes em um conjunto de dados têm diferentes níveis de importância.
- Cada valor (x) é multiplicado por seu peso e o total é dividido pela soma dos pesos (w).

$$\bar{x} = \frac{\sum_{i=1}^n w_i \bar{x}_i}{\sum_{i=1}^n w_i}$$

Média

- 5. Média Quadrática (Raiz Quadrática Média)
- Usada ao lidar com valores ao quadrado, como variância ou desvio padrão.

$$x_{\text{RMS}} = \sqrt{\frac{1}{n} (x_1^2 + x_2^2 + \cdots + x_n^2)}$$

- 6. Média Aparada
- Calculada removendo valores discrepantes (uma porcentagem especificada dos maiores e menores valores) do conjunto de dados antes do cálculo da média aritmética.
- Ajuda a reduzir o efeito de valores extremos.

Moda

- A última medida de centro abordada neste curso é a **moda**, que é a observação que ocorre com **mais frequência**.
- Se pelo menos duas observações ocorrem com mais frequência, o conjunto de dados tem múltiplas modas; se todas as observações ocorrem uma vez, não há moda.
- Por exemplo, $c = (3, 4, 3, 7, 7)$
- Vamos organizar de forma ascendente para a nossa análise.
- $c = (3, 3, 4, 7, 7)$
- Nesse, contexto os elementos "3" e "7" ocorrem com mais frequência. Ou seja, nossa série é bimodal (existem duas modas)!

Comandos [básicos] do Python para análises Estatística Rápida...

- `mean()` = média aritmética
- `geometric_mean()` = média geométrica
- `harmonic_mean()` = média harmônica
- `median()` = mediana
- `mode()` = moda
- Para demais comandos...
- <https://docs.python.org/3/library/statistics.html>

Média Vs Mediana

- Tanto a média quanto a mediana são **medidas centrais** de uma distribuição. Quando uma distribuição é **simétrica**, a média e a mediana são iguais. No entanto, é melhor usar a média para descrever a tendência central de uma distribuição simétrica.
- Todavia, quando uma **distribuição é assimétrica** ou contém valores discrepantes (*outliers*), é melhor usar a **mediana**. Isso ocorre porque a média inclui todas as observações de um conjunto de dados e, como tal, é facilmente influenciada por valores extremamente grandes ou pequenos (chamados outliers).
- Por outro lado, a mediana não inclui todas as observações de um conjunto de dados, mas apenas o(s) valor(es) mais central(is). Por esse motivo, a mediana é altamente resistente a valores discrepantes (outliers).

Média Vs Mediana

- Exemplo:
- Imagine que um indivíduo disciplinado coma de forma saudável e regrada de domingo à domingo.
- Um dia, este indivíduo decide comer alguns pedaços de pizza e beber uma cerveja em uma festa de aniversário com seus amigos.
- Esses pedaços de Pizza afetarão a ganha de peso do indivíduo ao longo prazo?
- Se calculamos a média de calorias consumidas [*diárias*] provavelmente o resultado seria muito alto.
- Todavia, se avaliarmos pela **mediana** a discrepância não é tão significativa assim....
- ***AVISO: Isso não é uma recomendação médica, apenas um exemplo imaginário para compreensão do exercício!!***

Revisão Medidas Centrais

- *Aqui estão algumas diretrizes para escolher a medida adequada para descrever o centro de uma distribuição:*
- Use a **mediana** quando a distribuição for **extremamente assimétrica** ou quando houver valores **discrepantes**.
- Use a **média** quando a distribuição for simétrica e **não houver valores discrepantes**.
- Para dados **qualitativos/categóricos**, podemos usar apenas a moda para descrever o centro.
- Para dados quantitativos, a **moda** também pode ser calculada. No entanto, ela **não é tão informativa** quanto a mediana ou a média.

Dispersão: IQR, Variância e Desvio Padrão

- Além do centro, precisamos de outra medida descritiva para descrever como os dados se **espalham**.
- Isso é chamado de dispersão ou variabilidade da distribuição. As medidas de variação abordadas são amplitude, amplitude interquartil (IQR do *inglês Inter-Quartile Range*) e desvio-padrão.

IQR

- Uma medida intuitiva da dispersão é o intervalo dos dados, que é definido como a diferença entre as maiores e as menores observações:
- $Intervalo = máximo - mínimo \equiv Q3 - Q1$
- Semelhante à média, a amplitude é sensível a valores discrepantes.

Variância

- Em termos gerais, ele fornece a distância média quadrada de cada observação x_i até a média da amostra \bar{x} .

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

- Em outros casos devemos calcular a **variância amostral**, no qual:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- Na 3ª aula explicaremos a diferença entre **variância populacional** e amostral.

Demonstração Operacional

x_i	x_i^2
3	$3^2=9$
5	$5^2=25$
3	$3^2=9$
7	$7^2=49$
7	$7^2=49$
$\sum x_i = 25$	$\sum x_i^2 = 141$

Fonte: Introduction to Applied Statistics (MacEwan University, 2024)

Demonstração Operacional

x_i	Deviation: $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	$3 - 5 = -2$	$(-2)^2 = 4$
5	$5 - 5 = 0$	$0^2 = 0$
3	$3 - 5 = -2$	$(-2)^2 = 4$
7	$7 - 5 = 2$	$2^2 = 4$
7	$7 - 5 = 2$	$2^2 = 4$
$\sum x = 25$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 16$

Fonte: Introduction to Applied Statistics (MacEwan University, 2024)

Desvio Padrão

- Assim como a média, o desvio padrão leva em consideração todas as observações e mede a variação, indicando, em média, **a distância** entre as observações e a média.
- Para um conjunto de dados com grande variação, ou seja, com observações muito diferentes entre si, o desvio padrão será grande.
- Para um conjunto de dados com pequena variação, em média, as observações estão próximas da média, portanto, o desvio padrão será pequeno.
- O calculo do Desvio Padrão é simples, é apenas a raiz quadrada da **variância**.

Desvio Padrão

- O calculo do Desvio Padrão é simples, é apenas a raiz quadrada da **variância**.

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{16}{5-1}} = 2$$

Assimetria

- Assimetria é uma maneira de medir a forma de uma distribuição: em particular, refere-se à assimetria que uma distribuição pode apresentar.
- A distribuição pode ser assimétrica para à direita ou à esquerda, dependendo de qual lado a *cauda* está.
- Dados **reais** provavelmente serão assimétricos de alguma forma. É fácil detectar assimetria em seus dados verificando se a média e a mediana do seu conjunto de dados não são aproximadamente iguais.

Assimetria

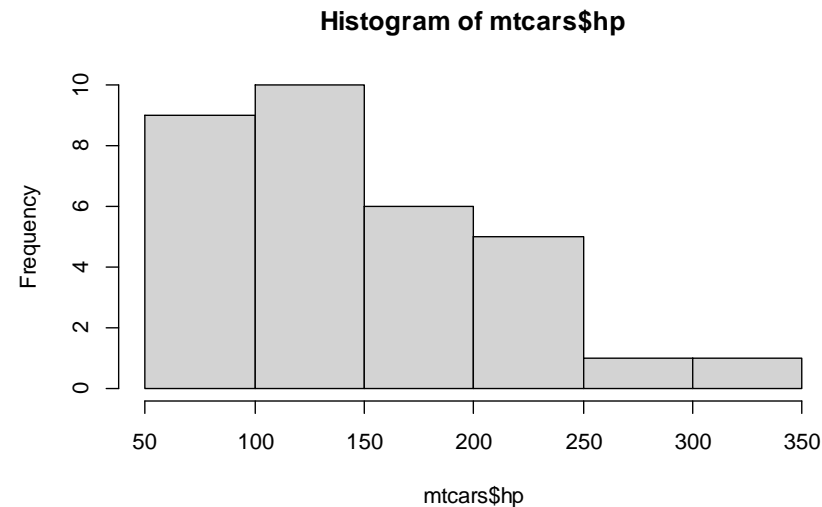
- Com dados assimétricos, o valor médio às vezes deixa de ser uma medida precisa da tendência central: em vez disso, a mediana é uma medida melhor, pois representa o valor central em um conjunto de dados e, portanto, descreve com mais precisão o centro da distribuição de dados.
- Existem várias maneiras de calcular Assimetria, mas, vamos utilizar aqui uma das mais simples a assimetria de Pearson.
- $Pearson = 3 \cdot \frac{média - mediana}{Desvio\ Padrão}$

Assimetria

- **Regrinhas de bolso:**
- Dados com +1 ou mais, ou -1 ou menos, são considerados altamente assimétricos positiva/negativamente.
- Uma assimetria positiva ou negativa mais moderada situa-se entre +0,5 e +1, ou -0,5 e -1, respectivamente.
- Uma assimetria positiva ou negativa muito leve situa-se entre 0 e +1, ou -0,5 e 0, respectivamente. Em testes de hipóteses, dados ligeiramente assimétricos podem ser tratados como normais, desde que apresentem simetria aproximada.
- Dados com assimetria **zero** são considerados **não assimétricos**.

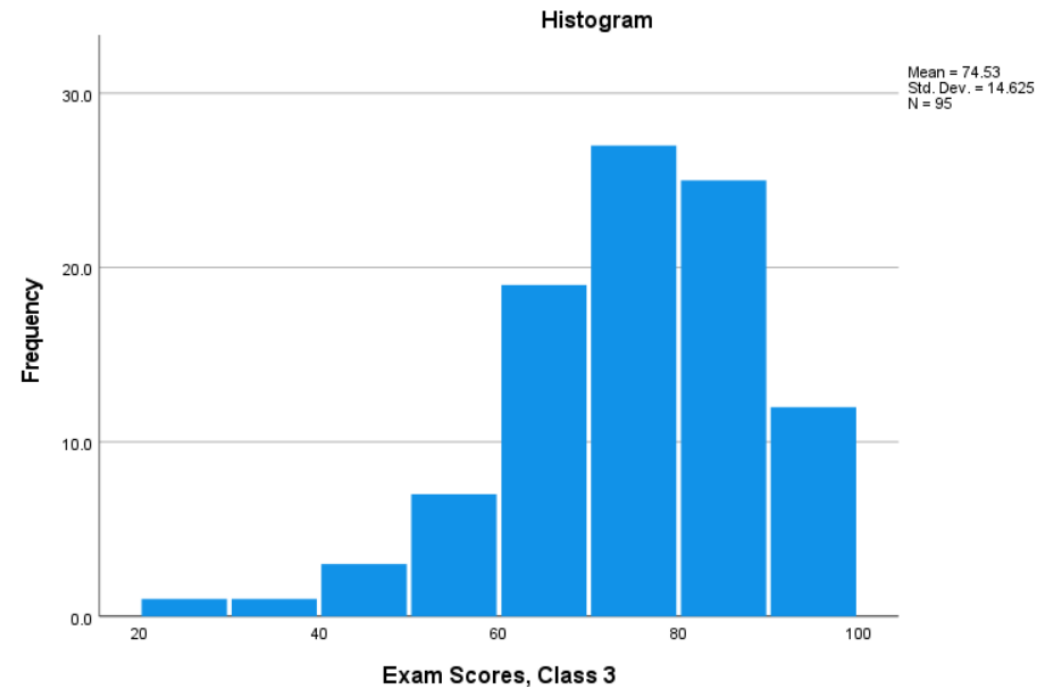
Assimetria [positiva] à direita

- A assimetria à direita ocorre quando o valor médio do conjunto de dados é maior que o valor mediano.
- Aqui, a "saliência" do histograma se situa à esquerda e **diminui para a direita** — trata-se de dados com distorção à direita e não são mais normais.
- Nesse caso, Hp (ou Poder da aceleração) tenderam a cair em seu desempenho, já que a maioria dos dados se situa na extremidade inferior da escala.



Assimetria [Negativa] à esquerda

- Em contraste, dados assimétricos à esquerda, também conhecidos como dados assimétricos negativos, são dados não simétricos e com **cauda para o lado esquerdo**.
- Se a média do conjunto de dados for menor que a mediana, o conjunto de dados terá uma assimetria à esquerda.
- Podemos dizer que mais alunos tiveram um bom desempenho nesta aula em comparação às duas anteriores, pois a concentração está no extremo superior da escala.



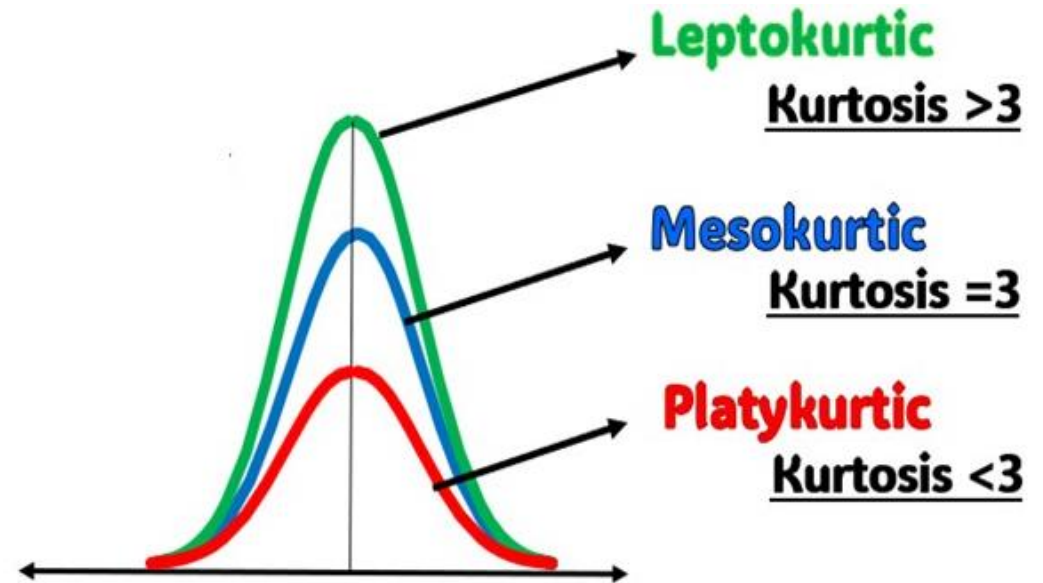
Fonte: University of South-Hampton
(2024)

Curtose

- Curtose é uma medida da cauda de uma distribuição — ou seja, a aparência das extremidades de uma curva de distribuição. A cauda indica a **presença de outliers**.
- As caudas podem, é claro, ser grossas, finas ou intermediárias: diríamos que uma *distribuição normal* tem cauda normal.
- Existem muitas fórmulas para calcular a curtose de uma distribuição, embora, novamente, possa ser mais desejável calculá-la usando um software em vez de uma fórmula.
- A curtose é frequentemente descrita em termos de "*curtose excessiva*", ou seja, curtose que é **diferente de um valor de 3**: isso ocorre porque a distribuição normal tem uma curtose de 3. A curtose excessiva é, portanto, calculada como:
- $\text{curtose excessiva} = 3 - \text{curtose}$

Curtose

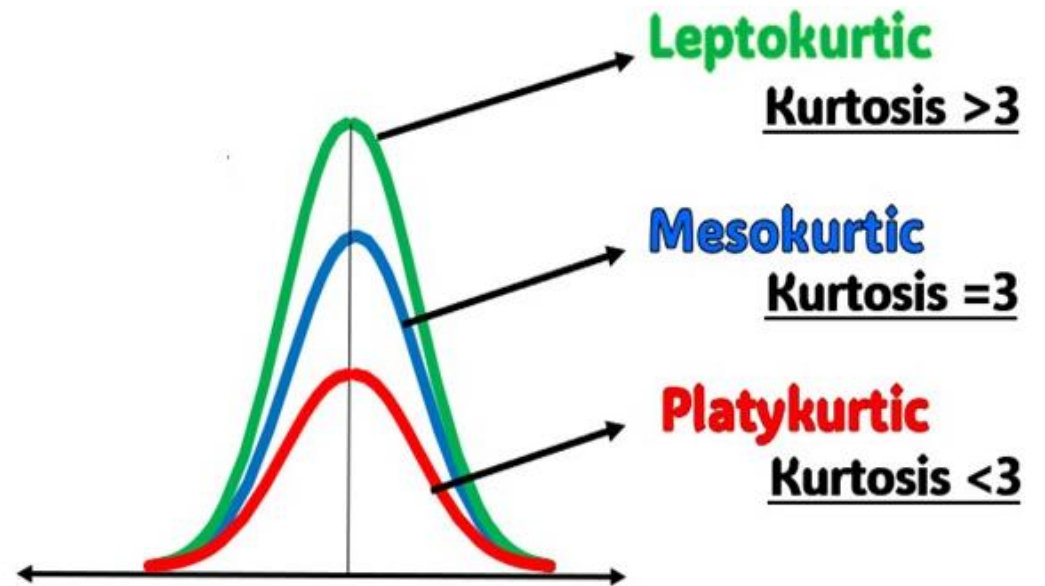
- Existem três tipos de curtose: *mesocurtose*, *platicurtose* e *leptocurtose*.



leptocurtose (verde); mesocurtose (azul); e, platicurtose (vermelho). **Fonte:** Six Sigma Worldwide (2024)

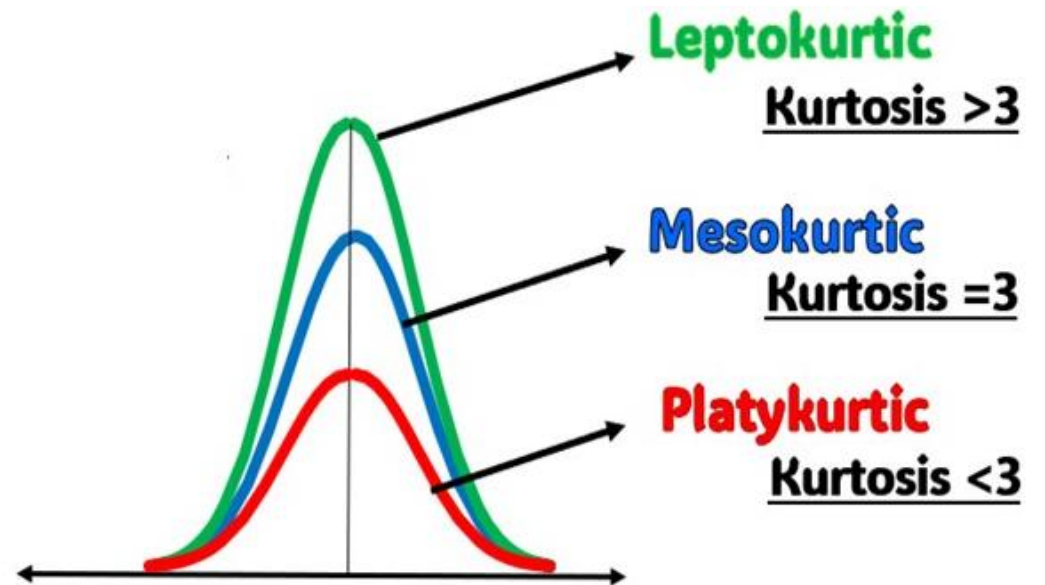
Curtose

- *Mesocurtose*
- Distribuições mesocúrticas são aquelas que seguem mais ou menos uma distribuição normal típica e são distribuições com cauda média e, portanto, não são consideradas como tendo cauda excessiva.
- Uma distribuição com curtose de aproximadamente 3 seria considerada mesocúrtica.



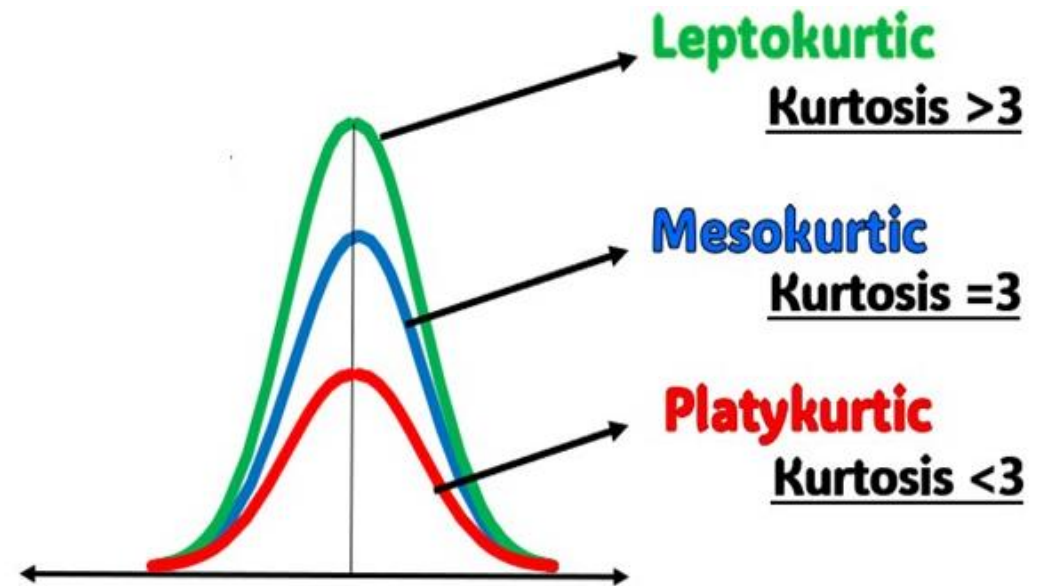
Curtose

- *Platicocurtose*
- Distribuições platicúrticas são aquelas com caudas finas. Uma distribuição com curtose menor que 3 (e, portanto, curtose excessiva negativa) é platicúrtica.



Curtose

- *Leptocurtose*
- Distribuições leptocúrticas são distribuições com caudas grossas. Distribuições com curtose maior que 3 (e, portanto, com curtose excessiva positiva) são leptocúrticas.
- Esses tipos de distribuições são mais propensos à presença de outliers, pois a maioria dos pontos próximos à média resulta em um desvio-padrão menor.



Exercícios para praticar...

- * Não precisa entregar...
- 1. Em 2004, o patrimônio líquido médio das famílias nos Estados Unidos era de US\$ 448,2 mil e o patrimônio líquido *mediano* era de US\$ 93,1 mil. Qual medida de centro você considera mais apropriada? Explique sua resposta.

Exercícios para praticar...

- * Não precisa entregar...
- 2. Wayne Gretzky, jogador profissional de hóquei aposentado, jogou 20 temporadas na National Hockey League (NHL), de 1980 a 1999. O número de jogos em que Gretzky jogou durante cada uma de suas 20 temporadas na NHL é o seguinte: 74, 80, 73, 78, 78, 45, 80, 79, 79, 80, 48, 64, 80, 70, 80, 74, 82, 81, 80, 82.
- A. Encontre a média, a mediana e a moda desses 20 números. Interprete as três medidas para o centro.
- B. Encontre os quartis dos dados e interprete.
- C. Encontre a amplitude, a amplitude interquartil e o desvio padrão amostral dos dados e interprete.
- D. Encontre o resumo de cinco números dos dados e desenhe um boxplot com esses 20 números. Comente o boxplot resultante. Escolha medidas adequadas para o centro e a dispersão (variação) da distribuição. Justifique sua resposta.
- **Fonte:** *Introduction to Applied Statistics (MacEwan University, 2024)*

Bibliografia

- **BUSSAB, W. O.; MORETTIN, P. A.** Estatística Básica. Saraiva, 2017.
- **LARSON, R.; FARBER, B.** Estatística Aplicada. Pearson, 2016.
- **Pishro-Nik; H.** Introduction to Probability, Statistics, and Random Processes. Kappa Research, 2014.
- **TRIOLA, M. F.** Introdução à Estatística. Pearson, 2018.