

# Temas para esse vídeo

---

## ▶ Temas

- ▶ Revisão: Tabelas de Frequência
- ▶ Medidas de dispersão

## ▶ Bibliografia básica

- ▶ Barrow, M. Estatística para economia, contabilidade e administração. São Paulo: Ática, 2007, Cap. 1
- ▶ Morettin, P. e W. Bussab. Estatística básica. 5. ed. São Paulo: Saraiva, 2005. Cap. 3

# Aula passada

---

- ▶ Quando queremos estudar a distribuição de valores que assume uma variável, podemos agrupar estes valores em intervalos
- ▶ A distribuição de frequência é um agrupamento de dados em classes, ou intervalos, para os quais se observa o número de observações em cada classe

# Aula passada

---

- ▶ Lembrando conceitos que vimos na aula passada:
  - ▶ Frequência do valor de uma variável é o número de repetições desse valor
  - ▶ Relacionando os valores que assume uma variável e suas frequências respectivas, temos a **distribuição de frequências absolutas**
  - ▶ **Frequência relativa** do valor de uma variável é obtida dividindo sua frequência absoluta pelo valor da amostra  
⇒ **distribuição de frequências relativas**
  - ▶ **Frequência acumulada** de uma variável é a soma das freq. absolutas e relativas desde o valor inicial da variável

# Aula passada

---

- ▶ Tabelas de frequência, gráficos e um ordenamento dos dados são instrumentos poderosos para resumir essas informações sobre o comportamento de uma variável
- ▶ Mas, muitas vezes, precisamos resumir de forma ainda mais concisa e encontrar um ou poucos valores que digam muito sobre uma série de dados, que sejam representativos dela

# Aula passada

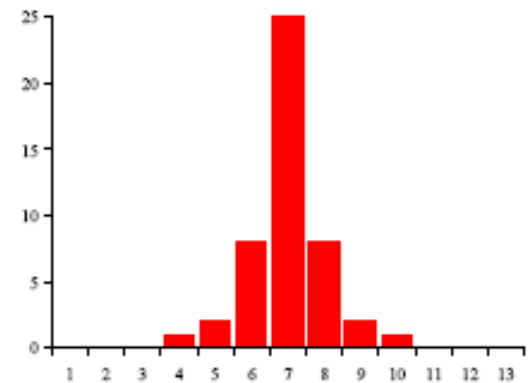
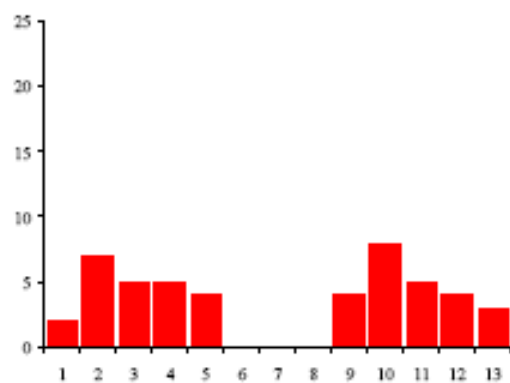
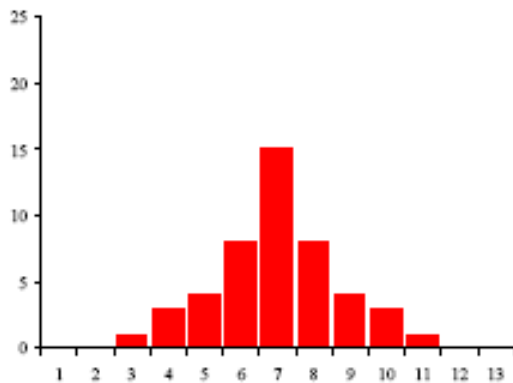
---

- ▶ Vimos as seguintes medidas de posição central:
  - ▶ **Moda:** é a realização mais frequente do conjunto de valores observados.
  - ▶ **Mediana:** é a realização que ocupa a posição central na série, quando os dados estão organizados em ordem crescente.
  - ▶ **Média aritmética:** como bem sabemos, é a soma dos valores observados dividida pelo número de observações.

# Medidas de dispersão

---

- ▶ A média ou a mediana nos dão indicações sobre o centro da distribuição
- ▶ Mas conhecer a posição central não nos diz muito acerca da variabilidade do conjunto das observações
- ▶ Por exemplo, as observações abaixo, com médias e medianas iguais a 7 mostram distribuições muito distintas



# Medidas de dispersão

---

- ▶ A **amplitude** é a medida da distância entre as observações máxima e mínima.
- ▶ **Amplitude** = valor máximo – valor mínimo
- ▶ Qual é a amplitude das seguintes notas da turma de MEPI?

Nota	Freq.	Freq. Acumulada
2	5	5
3	13	18
4	29	47
5	33	80
6	17	97
7	8	105
8	4	109
9	1	110

# Medidas de dispersão

---

- ▶ Claramente, a amplitude está fundamentada em duas observações extremas
- ▶ Para verificar melhor como os valores de uma variável variam em torno da média é preciso utilizar outras medidas
- ▶ O intervalo entre quartis é definido como a diferença entre o terceiro e o primeiro quartil



# Pensando no ordenamento dos dados

Medidas relacionadas a diversas partes de um conjunto de dados são úteis na apresentação da distribuição de seus valores, principalmente se o conjunto de dados é não simétrico.

Quartis - 1º e 3º Quartis (25% e 75%)



# Quantis

---


- ▶ Por exemplo, se tivermos uma série de dados referentes a uma variável, como as notas obtidas por uma classe de 9 alunos

2, 4, 8, 9, 5, 10, 7

Ordenando os valores, teríamos

$2 < 4 < 5 < 7 < 8 < 9 < 10$

A mediana seria  $q(0,50)=7$

- ▶ Aqui teríamos que usar os conceitos que aprendemos vendo as distribuições de frequência e criar intervalos aos quais relacionaríamos o número de observações encontradas em cada faixa de valor
- 
- 

# Medidas de dispersão

---

- ▶ Considere o nosso exemplo anterior. Qual é o intervalo entre quartis?

Nota	Freq.	Freq. Acumulada
2	5	5
3	13	18
4	29	47
5	33	80
6	17	97
7	8	105
8	4	109
9	1	110

# Medidas de dispersão

---

- ▶ O intervalo entre quartis nos dá os valores entre os quais estão 50% das observações.
- ▶ Dessa forma, ele considera a dispersão interna da distribuição
- ▶ Porém, não utiliza toda a informação disponível
- ▶ Uma medida de dispersão que utiliza toda a informação disponível é aquela que mede a dispersão dos dados em torno da sua média. Utilizam-se principalmente duas medidas
  - ▶ Desvio médio
  - ▶ Variância e respectivo desvio-padrão

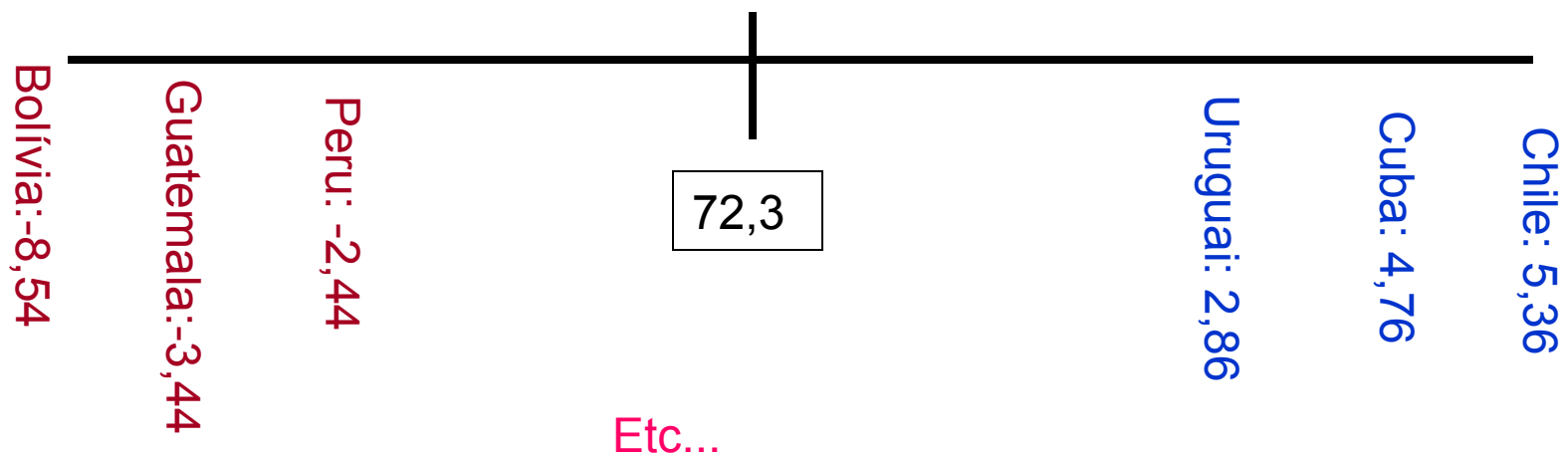
# Considere a expectativa de vida em alguns países da América Latina

Países	Expect (anos)
Argentina	74.30
Bolivia	63.80
Brazil	71.00
Chile	77.70
Colombia	71.60
Cuba	77.10
Ecuador	74.20
El Salvador	70.60
Guatemala	68.90
Mexico	74.80
Paraguay	70.80
Peru	69.90
Uruguay	75.20
Venezuela	72.80



# Variância

- ▶ Consideremos, de novo, o exemplo da expectativa de vida. O valor médio é 72,3
- ▶ Se somarmos todas as diferenças, teremos zero. Por isso, tomamos o quadrado ou o valor absoluto das diferenças



# Medindo os desvios

---

- ▶ Desvio médio

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

- ▶ Variância

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

- ▶ Teríamos, para nosso exemplo

- ▶  $dm(X) = 2,82$
- ▶  $\text{var}(X) = 13,95$



# Medindo os desvios

---

- ▶ A variância tem dimensão igual ao quadrado da dimensão dos dados originais. No nosso caso, isso pode parecer curioso e de difícil interpretação (anos ao quadrado)
- ▶ Usa-se então a raiz quadrada da variância, que é o desvio padrão
- ▶ O desvio padrão é a raiz quadrada positiva da variância

$$dp(X) = \sqrt{\text{var}(X)}$$





# Calculando para a expectativa de vida dos 14 países da América Latina

---

Países	Expect (anos)	Desvios	Desvios absolutos
Argentina	74.30	1.96	1.96
Bolivia	63.80	-8.54	8.54
Brazil	71.00	-1.34	1.34
Chile	77.70	5.36	5.36
Colombia	71.60	-0.74	0.74
Cuba	77.10	4.76	4.76
Ecuador	74.20	1.86	1.86
El Salvador	70.60	-1.74	1.74
Guatemala	68.90	-3.44	3.44
Mexico	74.80	2.46	2.46
Paraguay	70.80	-1.54	1.54
Peru	69.90	-2.44	2.44
Uruguay	75.20	2.86	2.86
Venezuela	72.80	0.46	0.46

Média	72.3
Mediana	72.2
Desvio médio (desv.medio(..))	2.8
Variância (soma desvios) <sup>2</sup> /N	12.3
Desvio padrão	3.6



# As medidas de dispersão vistas

---

- ▶ Desvio médio

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

- ▶ Variância

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

- ▶ Desvio padrão

$$dp(X) = \sqrt{\text{var}(X)}$$

---



# Retomando as medidas já vistas

---

- ▶ Até agora, consideramos sempre as medidas para a população
- ▶ Para a amostra, as fórmulas ficam:

▶ Variância da amostra  $\Rightarrow$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

▶ O desvio padrão da população era  $\Rightarrow$

$$\sigma_X = +\sqrt{\sigma_X^2}$$

▶ Para a amostra é  $\Rightarrow$

$$S_X = +\sqrt{S_X^2}$$



# Resultado da variância tem características importantes

---

- ▶ Variância é sempre um número positivo
- ▶ O numerador, quando se calcula a variância para a população ou uma amostra é o mesmo (soma dos desvios ao quadrado)
- ▶ A variância de uma variável para a população é uma média aritmética dos quadrados dos desvios
- ▶ A variância de uma variável para a amostra é uma média, embora seja dividida por  $n-1$
- ▶ A variância é afetada por valores extremos



# Desvio padrão

---

- ▶ A variância é medida em termos do quadrado da unidade de medida original da amostra.
- ▶ Como vimos, o desvio padrão é  $dp(X) = \sqrt{\text{var}(X)}$
- ▶ Portanto, o resultado está expresso em uma unidade de medida que faz algum sentido
- ▶ Alto desvio padrão indica uma maior variação dos dados em relação à média

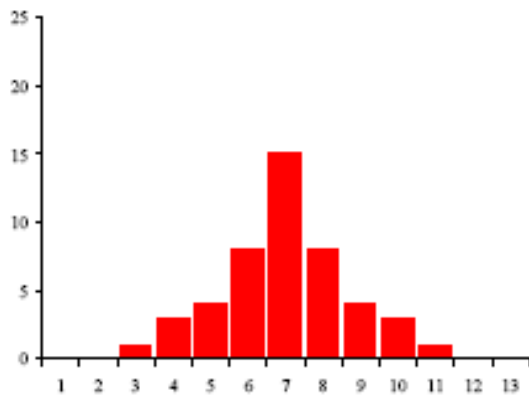


# Exemplos de desvio padrão

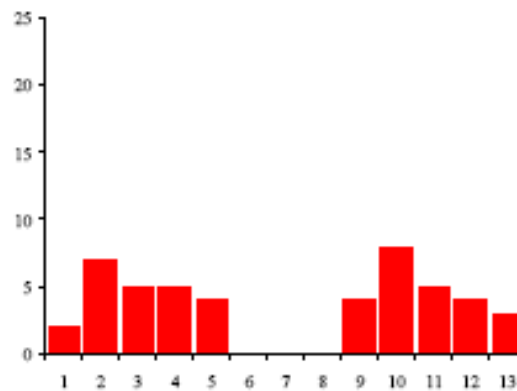
A distribuição dos valores em relação à média pode ser muito distinta.

Dizemos que os valores que assumem a variável dão origem a diferentes distribuições

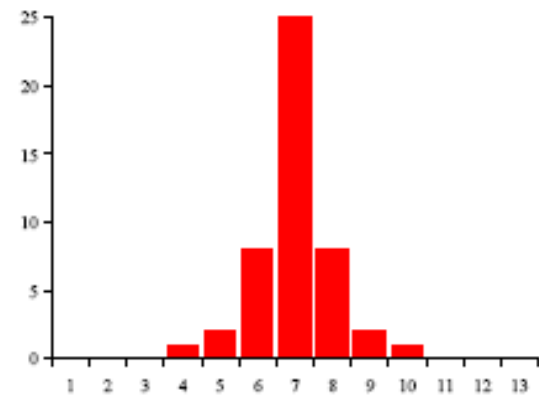
Duas variáveis com médias iguais e desvio padrão diferentes têm diferentes formas de distribuição de frequência



Distribuição a  
 $S = 1,69$



Distribuição b  
 $S = 4,03$



Distribuição c  
 $S = 1,04$

# Desvio padrão

---

- ▶ Só são comparáveis desvios padrão que estejam na mesma unidade de medida
- ▶ O desvio padrão será 0 quando todas as observações forem iguais
- ▶ Quanto maior a variação, maior o desvio padrão
- ▶ Uma boa forma de pensar em  $S$  é como uma distância média entre uma observação e a média



# Coeficiente de variação: medida relativa de dispersão

---

- ▶ O desvio-padrão é uma medida de dispersão absoluta
- ▶ Nem sempre podemos comparar medidas de dispersão de duas ou mais distribuições: as unidades podem ser diferentes
- ▶ O **coeficiente de variação** é uma medida relativa de dispersão, que permite comparar distribuições

$$CV_{pop} = \frac{\sigma_X}{\mu_X}$$

$$CV_{amo} = \frac{S_X}{\bar{X}}$$

⇒ A variável com menor CV tem menor dispersão ou variabilidade

---





# Algumas falhas

---

- ▶ Tanto média como o desvio padrão podem apresentar falhas para representar um conjunto de dados, já que eles
  - ▶ Não nos dão, por exemplo, uma idéia da simetria ou assimetria da distribuição de dados



# Medida de assimetria

---

- ▶ Além da dispersão das observações em um conjunto de dados, podemos saber como se distribuem esses dados, se estão mais dispersos “para um lado ou para outro”
- ▶ O coeficiente de assimetria mostra isso
  - ▶ Se  $CA > 0$  distribuição tem assimetria à direita
  - ▶ Se  $CA < 0$  distribuição tem assimetria à esquerda
  - ▶ Se  $CA = 0$  distribuição é simétrica



# Medida de assimetria

---

- ▶ Coeficiente de assimetria

$$CA = \frac{\sum (x - \mu)^3}{N\sigma^3}$$

$$CA = \frac{\sum (x - \bar{X})^3}{NS^3}$$

- ▶ Para ter uma idéia ainda melhor da distribuição dos dados ao redor da média, é sempre bom fazer um gráfico de frequência.

