DISTRIBUIÇÃO DE FREQUÊNCIA
MEDIDAS DE TENDÊNCIA CENTRAL
MEDIDAS DE DISPERSÃO

Estatística Aplicada I
IRI-USP

Maio 2021

Profª. Maria Antonieta Del Tedesco Lins

1

Aula de hoje

2

- Temas
  - Distribuições de frequência
  - Medidas de tendência (posição)
  - Medidas de dispersão

- Bibliografia básica
  - Agresti, A. e Finlay, B. Métodos Estatísticos para as Ciências Sociais. 4ªed. Porto Alegre: Penso, 2012 Cap.2
  - Barrow, M. Estatística para economia, contabilidade e administração. São Paulo: Ática, 2007, Cap. 1
  - Lapponi, J. Estatística usando Excel 5 e 7. São Paulo: Lapponi Treinamento e Editora, 1997. Capítulos 3 e 4
  - Morettin, P. e W. Bussab. Estatística básica. 5. ed. São Paulo: Saraiva, 2005. Cap. 3

2

3 Distribuições de frequência

3

Tabelas de frequência

4

- 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

4

Tabelas de frequência e dados quantitativos discretos

5

- Lembrando conceitos que vimos em aulas anteriores
  - 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

5

Exemplo: Vamos considerar um conjunto de observações desordenadas

Faixa etária de crianças participando de um acampamento

6	10	9	14	7	4
8	11	12	5	9	13
9	10	8	6	7	14
11	6	12	11	15	13
12	11	4	10	7	13
10	9	8	12	13	7

Antes de tudo, é difícil ver como se concentram as idades das crianças ou, ainda, qual é a faixa etária dos participantes

6

Deveríamos, então, ordenar as informações

4	6	8	10	11	13
4	7	8	10	12	13
4	7	8	10	12	13
5	7	9	10	12	14
6	7	9	11	12	14
6	8	9	11	13	15

- A esta ordenação, chama-se rol

7

Depois, fica fácil estabelecer a frequência

Idade	Frequência
4	3
5	1
6	3
7	4
8	4
9	4
10	4
11	3
12	4
13	4
14	2
15	1

8

Elementos de uma distribuição de frequência

9

- Classes: caso as colunas da tabela de distribuição de frequência contenham muitos valores elencados, podemos reduzir a quantidade desses valores elencados agrupando-os em intervalos.
- Esses agrupamentos de valores num intervalo de abrangência são chamados de **classes**

9

Para nosso exemplo, as classes ficam

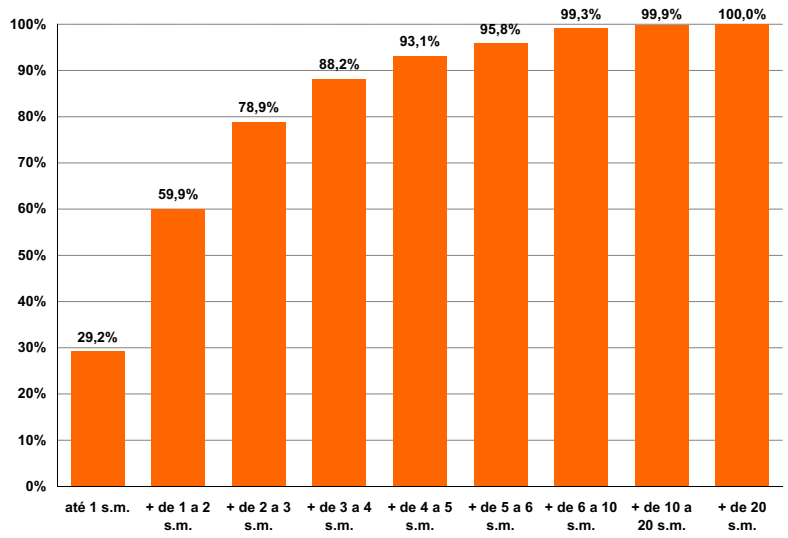
Idade	Frequência
4l-6	4
6l-8	7
8l-10	8
10l-12	7
12l-14	8
14l-16	3

10

10

Exemplo: Distribuição acumulada do déficit habitacional por inadequação, por faixa de renda, Brasil, 2006

11



Fonte: FGV Projetos

11

Tabelas de frequência com dados contínuos

12

- Quando não se trabalha com valores inteiros (variáveis discretas), fica inviável determinar o número de vezes que um valor ocorre
- Por isso, o interessante é trabalhar com classes de valores
  - Definir a quantidade, limites e amplitude das classes
  - Muitas vezes, a escolha de intervalos e número de classes pode ser arbitrária e depender da sensibilidade do pesquisador
  - Com um pequeno número de classes, pode-se perder informação e um número grande dificulta o resumo dos dados

12

Construção da tabela de frequência

- Algumas orientações práticas
  - Número de classes: Não existe uma regra única para a definição do número de classes
  - Muitas vezes, vale a percepção do pesquisador
  - Algumas diretrizes podem servir
  - Para uma amostra de tamanho n, a quantidade k de classes recomendadas pode ser
    - $k = \sqrt{n}$ arredondando o resultado inteiro para menor ou maior
    - k = 1 + 3,322 x log(n), arredondando o valor inteiro
    - Ainda se pode usar k como o menor valor inteiro que satisfaça a condição $2^k \leq n$ (o mesmo que a primeira fórmula com $2^k = n$)
  - Os valores da variável a serem inseridos em classes se tornam em uma nova variável definida pelos limites dos intervalos

13

Construção da tabela de frequência

14

- Como dito, vale ir experimentando o número de classes de forma a encontrar uma distribuição que represente bem os valores de uma variável
- Quando se trabalha com classes, a tabela de frequências perde a identidade de cada observação ⇨ ocorre perda de informação
- Os valores da variável transformam-se em uma nova variável cujos valores são os limites dos intervalos determinados

14

Usando o exercício feito em Lapponi

15

Comparação dos métodos sugeridos para a escolha da quantidade de classes

Tamanho da amostra n	Quantidade de classes		
	$k=n^{0,5}$	$k=1+3,322\times\log(n)$	$k=\log(n)/\log(2)$
10	3.16	4.32	4
20	4.47	5.32	5
30	5.48	5.91	5
40	6.32	6.32	6
50	7.07	6.64	6
60	7.75	6.91	6
70	8.37	7.13	7
80	8.94	7.32	7
90	9.49	7.49	7
100	10.00	7.64	7
150	12.25	8.23	8
200	14.14	8.64	8
250	15.81	8.97	8
300	17.32	9.23	9
350	18.71	9.45	9
400	20.00	9.64	9
450	21.21	9.81	9
500	22.36	9.97	9
750	27.39	10.55	10
1,000	31.62	10.97	10

Determinação de *k* para um *n* qualquer			
35	5.92	6.13	6

Fonte: Lapponi, Cap. 2

15

Exemplo 2.8 Lapponi, p. 45

- Objetivo é construir a tabela de frequências absolutas e relativas das vendas de uma empresa levantadas na tabela ao lado
- Quantidade de classes: ideal é que tenham todas a mesma amplitude
- Aplicando as fórmulas na tabela anterior, o número de classes (k) será 5 (n=25)
- Valores máximo e mínimo são 430 e 280
- Intervalo de variação é 150

Amostra
280
305
320
330
310
340
330
341
369
355
370
360
370
365
280
375
380
400
371
390
400
370
401
420
430

16

Exemplo 2.8 Lapponi, p. 45

- Amplitude as 5 classes é dada por

$$\frac{430-280}{5}=30$$

Limites	
Inferior	Superior
280	310
310	340
340	370
370	400
400	430

- Assim, constrói-se uma tabela de seleção

17

Exemplo 2.8/2.9 Lapponi, p. 45

- Fazendo as seleções dos valores entre as cinco classes, temos a dist. de freq ao lado

Limites		Tabela de Freqüencias			
Inferior	Superior	Absolutas	Relativas	Acumul. Abs.	Acumul. Rel.
280	310	3	12.00%	3	12.00%
310	340	4	16.00%	7	28.00%
340	370	6	24.00%	13	52.00%
370	400	7	28.00%	20	80.00%
400	430	5	20.00%	25	100.00%

- Para fazer o exercício com o excel é preciso ajustes nas classes (subtraiu-se 0.1 ao limite superior)

Limites		
Tec. Inferior	Tec. Superior	Excel
280	310	309.9
310	340	339.9
340	370	369.9
370	400	399.9
400	430	430

Limites	Tabela de Freqüencias			
Excel	Absolutas	Acumul. Abs.	Relativas	Acumul. Rel.
309.9	3	3	12.00%	12.00%
339.9	4	7	16.00%	28.00%
369.9	6	13	24.00%	52.00%
399.9	7	20	28.00%	80.00%
430	5	25	20.00%	100.00%
	0			

18

Medidas de tendência central

ou de posição

19

O que são essas medidas?

20

- 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
- Resumir uma série de dados significa reduzí-la de forma substancial
- Usamos medidas de ordenamento ou de posição quando queremos resumir e analisar uma amostra ou a população toda

20

Medidas de posição central

21

- Em geral, utiliza-se três medidas principais
  - **Moda**: é a realização mais frequente do conjunto de valores observados. No ex 2.8 de Lapponi visto há pouco, nos valores de vendas da empresa, a moda é 370, ou seja, é o valor de vendas que mais vezes aparece na amostra
  - **Mediana**: é a realização que ocupa a posição central na série, quando os dados estão organizados em ordem crescente. Naquele exemplo, a mediana é 369
  - **Média aritmética**: como bem sabemos, é a soma dos valores observados dividida pelo número de observações. Ex: só como curiosidade, calculamos que o preço médio da tonelada métrica de arroz entre janeiro de 1957 e novembro de 2008 foi de US$ 255,51

21

Formalizando os conceitos

- Se x_1, x_2, ..., x_n são os valores da variável X, a média aritmética pode ser escrita

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Se tivermos n observações de X, das quais n_1 são iguais a x_1, n_2 são iguais a x_2, a média pode ser escrita

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$$

  - Se $f_i = n_i/n$ for a frequência relativa da observação x_i, então

$$\bar{x} = \sum_{i=1}^{k} f_i x_i$$

22

Formalizando os conceitos

- Considerando as observações ordenadas em ordem crescente e sendo a menor observação $x_{(1)}$, etc., até $x_{(n)}$
- Assim: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$
- Ordenadas são chamadas estatísticas de ordem
- E a mediana pode ser defina por

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se n ímpar} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{se n par} \end{cases}$$

23

Propriedades da média

1. A soma dos desvios de uma amostra ou variável é sempre igual a zero
2. A soma dos quadrados dos desvios com relação à própria média é sempre um valor mínimo
   - Isso significa que, se os desvios fossem calculados com relação a qualquer outro valor diferente da média da amostra, a soma dos quadrados dos desvios seria um número maior

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right) = 0$$

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

⇩

mínimo

24

Análise do resultado da média

25

- Todos os valores da variável são incluídos no cálculo da média
- A média é um valor único
- A média está posicionada de forma equilibrada entre os valores ordenados da amostra. Ou seja, os valores se distribuem em torno da média
- A média é uma medida sensível à presença de dados extremos ou suspeitos
- Nas amostras ou variáveis com histograma simétrico, os valores da mediana, da moda e da média são iguais

25

Exemplo: Expectativa de vida ao nascer de 14 países latino-americanos, 2000- 2005

26

Países	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 (Bolivarian Republic of)	72,80

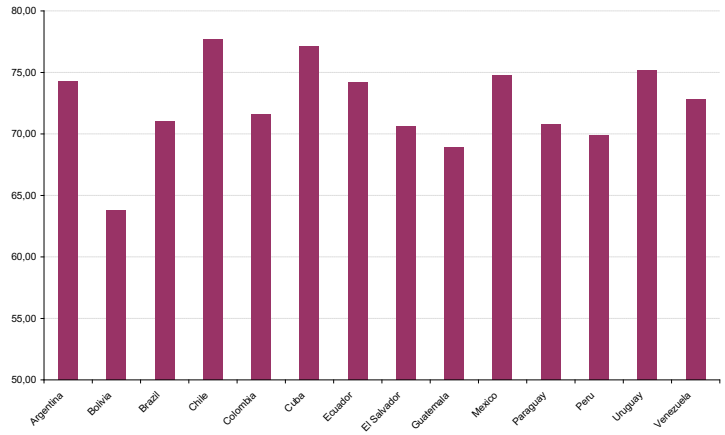
Fonte: CEPAL

- A expectativa de vida média destes países é 72,3 anos
- A mediana da expectativa de vida é 72,2 anos

26

No gráfico, vemos

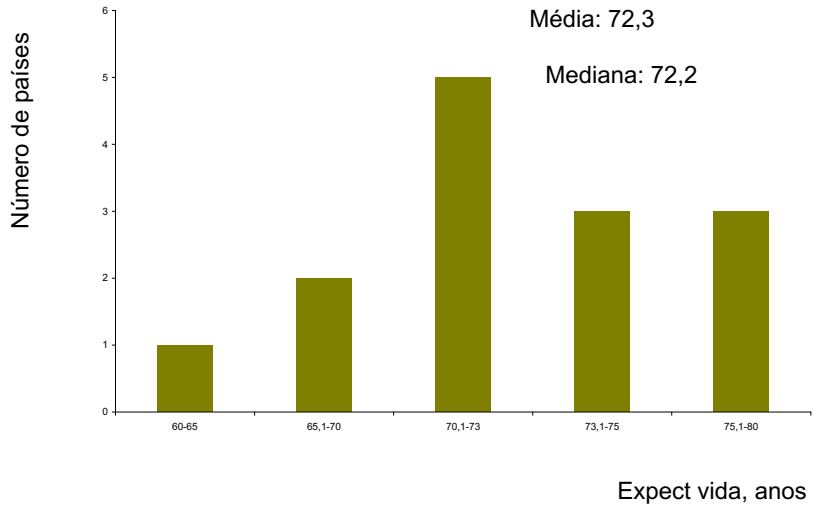
27



27

Distribuição de frequência

28



28

Análise das medidas de tendência central

- Moda

👍	👎
Fácil de calcular	Pode estar longe do centro dos dados
Não é afetada pelos extremos da amostra	Difícil de incluir em funções matemáticas
Pode ser aplicada em qualquer escala	Não usa todos os dados da amostra
	Pode haver mais de uma moda
	Certas amostras podem não ter nenhuma

29

Análise das medidas de tendência central

- Mediana

👍	👎
Fácil de calcular	Difícil de incluir em funções matemáticas
É um valor único	Não usa todos os dados da amostra
Pode ser aplicada em qquer escala	
Não é afetada pelos extremos da amostra	

30

Análise das medidas de tendência central

- Média

Fácil de compreender e aplicar	É afetada pelos extremos da amostra
Usa todos os dados da amostra	É necessário conhecer todos os dados da amostra
Pode ser aplicada nas escalas intervalar e proporcional	
É um valor único	
Fácil de incluir em funções matemáticas	

31

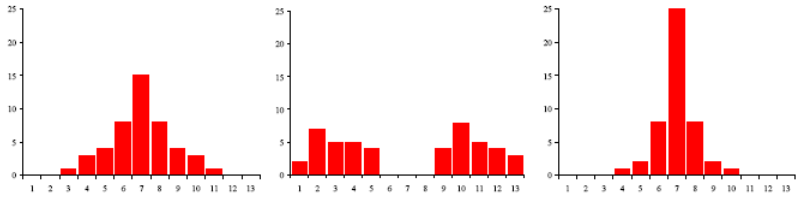
Medidas de dispersão

32

Medidas de dispersão

33

- 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 e 48 observações mostram distribuições muito distintas



33

Medidas de dispersão

34

- O intervalo é a medida da distância entre o valor maior e o menor
- O intervalo no nosso exemplo de expectativa de vida na Am. Latina é 77,7-63,8 = 13,9
- Claramente, essa não é uma informação muito precisa, pois ela está fundamentada em duas observações
- Para verificar melhor como os valores de uma variável variam em torno da média é usar outras medidas
- A ideia básica é medir quanto os valores individuais se afastam do valor médio
- Alguns destes desvios serão positivos, outros negativos, por isso devemos considerar o quadrado dos desvios

34

Medidas de dispersão

35

- Precisamos de medidas que sumarizem a variabilidade de um conjunto de observações e que nos permitam realizar comparações conjuntos diferentes de valores
- Um critério usado para efetuar essa comparação é o que mede a dispersão dos dados em torno da sua média. Utilizam-se principalmente duas medidas
  - Desvio médio
  - Variância

35

Desvio médio

- Retomando o exemplo da expectativa de vida em países selecionados da América Latina, teríamos os seguintes desvios ($x_i - \bar{x}$)

1,96	-8,54	-1,34	5,36	-0,74	4,76	1,86	-1,74	-3,44	2,46	-1,54	-2,44	2,86	0,46

- Observamos que a soma deles é igual a zero

$$\sum_{i=1}^{14}\left(x_i - \bar{x}\right) = 0$$

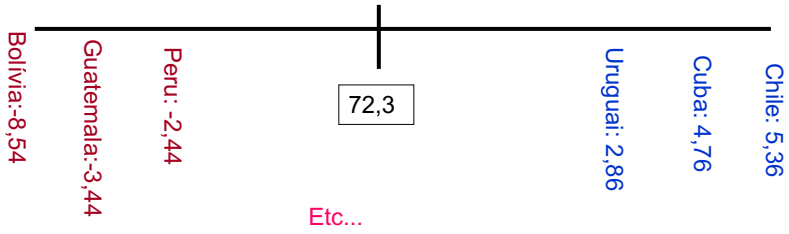
- Então, a soma dos desvios não é uma boa medida para a dispersão destes dados. Duas alternativas, exprimindo as medidas como médias
  - Desvio médio
  - Variância

36

Variância

37

- 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 das diferenças



37

Medindo os desvios

- Desvio médio

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

- Variância

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

- Teríamos, para nosso exemplo
  - dm(X)=2,82
  - var(X) = 13,95

38

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 o quadrado da variância, que é o desvio padrão
- O desvio padrão é a raiz quadrada positiva da variância

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

- Ambas medidas de dispersão indicam, em média, qual será o erro (desvio) incorrido ao tentar substituir cada observação pela média do conjunto de dados

39

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)^2/N	12.3
Desvio padrão	3.6

40

Outro exemplo

41

Bósnia

	Número de refugiados que deixaram o país (x1000)
1992	0
1993	0
1994	863
1995	906
1996	1006
1997	557
1998	424
1999	300
2000	250
2001	210
2002	160
2003	142
2004	30
2005	28
2006	30
2007	30

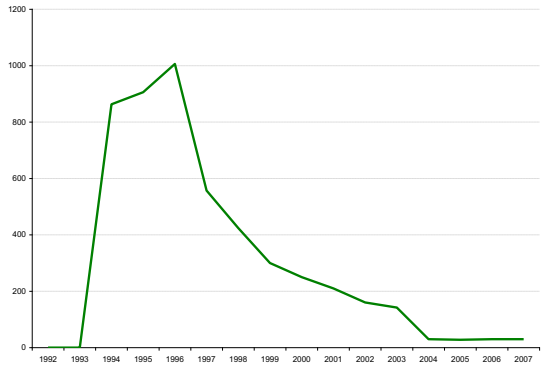
A leitura e interpretação desses dados requer informações históricas precisas, mas podemos resumi-los usando as técnicas que conhecemos

Fonte: United States Committee for Refugees and Immigrants (USCRI), World Refugee Survey (Annual Series). Disponível em www.systemicpeace.org

41

Antes de tudo, vamos ver o gráfico

Número de refugiados que deixaram a Bósnia (em milhares)



Fonte: United States Committee for Refugees and Immigrants (USCRI), World Refugee Survey (Annual Series). Disponível em www.systemicpeace.org

42

Um exemplo

- Usando nossos dados, temos

Amostra	Desvio	Desvio absoluto	Desvio^2
0	-309	309	95172
0	-309	309	95172
863	555	555	307470
906	598	598	357006
1006	698	698	486506
557	249	249	61752
424	116	116	13340
300	-9	9	72
250	-59	59	3422
210	-99	99	9702
160	-149	149	22052
142	-166	166	27656
30	-279	279	77562
28	-281	281	78792
30	-279	279	77562
30	-279	279	77562

- Fazendo os cálculos:

N	16
Média	309
Soma dos desvios absolutos	4427
Desvio absoluto medio	277
Desvio médio (desv.medio(..))	277
Soma desvios^2	1790804
Fórmula excel (=somaquad(desvios))	1790804
Variância (soma desvios)^2/N	111925
Fórmula excel (=varp(..))	111925

- Como descrever o fenômeno ilustrado pelos números? (mesmo fazendo pouco sentido histórico...)

43