

Amostragem Aleatória e Descrição de Dados - parte I

Marcos Oliveira Prates

2012/02

- 1 Amostra e População
- 2 Estatísticas Descritivas
- 3 Medidas de Posição
- 4 Medidas de Variabilidade

Objetivos

Ao final deste capítulo você deve ser capaz de:

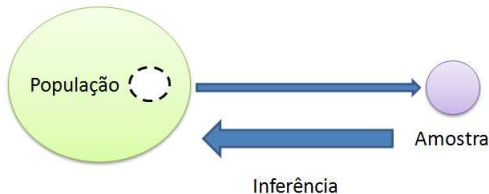
- Calcular e interpretar as seguintes medidas de uma amostra:
 - média;
 - mediana;
 - variância;
 - desvio padrão;
 - amplitude.

Introdução

- Queremos analisar uma variável:
 - anos de estudo dos professores universitários brasileiros.
- Geralmente não é possível estudar todo o conjunto de dados de interesse.
- Seja por questões de tempo ou custo.
- Seleccionamos uma parte desse conjunto: **amostra**.
- Os nossos dados de interesse constituem a **população**.

Amostra aleatória

- Selecionamos a amostra para obter informação sobre a população.
- Se queremos investigar as alturas dos alunos:
 - selecionamos alguns deles aleatoriamente e medimos suas alturas.
- Esse tipo de procedimento é chamado **amostragem aleatória**.
- São desenvolvidos modelos para descrever esse tipo de situação.



- A seleção da amostra pode ser feita de várias maneiras.
- O seu tamanho vai ser determinado a partir do erro máximo permitido.
- E também pelas características da população.
- Depende da variabilidade:
 - grande heterogeneidade $\Rightarrow$ tamanho de amostra maior.
- Também do custo de coleta.
- A amostra deve representar a população da melhor forma possível.

Outros tipos de amostragem:

Amostragem sistemática

- Um elemento é escolhido a cada k elementos.
- Não é recomendada quando os dados têm sazonalidade.

Amostragem Estratificada

- Divide a população em estratos de elementos que dividem uma característica.
- Seleciona-se uma amostra em cada grupo.
- Melhora a precisão das estimativas.
- Controla fatores que podem influenciar nas respostas dos elementos.
- Exemplo: divide-se a população por faixa etária.

Amostragem por conglomerado

- Divide a população em unidades: conglomerados.
- Seleciona-se alguns deles.
- Todos elementos do conglomerado vão fazer parte da amostra.
- Conglomerados podem reduzir o custo de coleta.
- A variabilidade dentro do conglomerado $>$ variabilidade entre os conglomerados.
- Exemplo: seleção de lotes de produtos em uma fábrica.

Descobrimos o óbvio?



- o que você vê nessa foto?
- Simplesmente um conjunto de grãos de café?

Descobrimos o óbvio?



- o que você vê nessa foto?
- Simplesmente um conjunto de grãos de café?
- Existe a face de um homem entre os grãos.

- O olho humano tem dificuldade de identificar padrões ou falta de padrão.
- Precisamos de medidas para resumir a informação.
- Usamos as **estatísticas descritivas**.

Exemplo:

- Em 1969 foi realizado um estudo em um determinado país.
- 7683 indivíduos foram pesquisados.
- Foram medidas as seguintes variáveis:
 - nível educacional;
 - peso (kg);
 - altura (cm);
 - idade (anos)
 - glicemia (mg/dL);
 - colesterol sérico (mg/dL);
 - pressão sistólica (mmHg).

ID	EDUCATIONA L LEVEL	WEIGHT (KG)	HEIGHT (CM)	AGE	SMOKING STATUS
1	2	70	165	61	1
2	1	60	162	52	0
3	1	62	150	52	1
4	2	66	165	51	1
5	2	70	162	51	0
6	4	59	165	53	0
7	1	47	160	61	0
8	3	66	170	48	1
9	5	56	155	54	0
10	2	62	167	48	0
11	4	68	165	49	1
12	1	65	166	48	0
13	1	56	157	55	0

- Sumários e apresentação de dados são essenciais para bom julgamento estatístico.
- Podemos assim focar nas características importantes dos dados.
- É útil descrever **numericamente** as características dos dados.
- Podemos caracterizar a localização ou tendência central:
 - média amostral.

Medidas de Posição

Média da Amostra

- Uma amostra de tamanho n é coletada.
- As observações dessa amostra são denotada por

$$x_1, x_2, \dots, x_n .$$

- A média amostral será

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} .$$

Exemplo:

- Um engenheiro está projetando um conector de náilon.
- Ele estabelece que o conector deve ter 3/32 polegadas.
- Mas não sabe se essa espessura fornecerá uma boa força de remoção.
- Ele faz oito unidades do protótipo e mede a força delas:

$$x_1 = 12,6 \quad x_2 = 12,9 \quad x_3 = 13,4 \quad x_4 = 12,3 \quad x_5 = 13,6$$

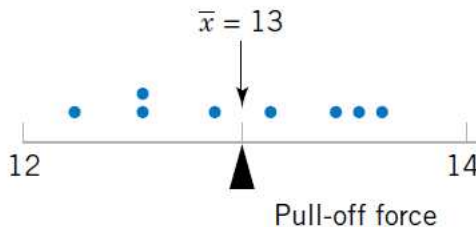
$$x_6 = 13,5 \quad x_7 = 12,6 \quad x_8 = 13,1 .$$

- A média amostral é

$$\bar{x} = \frac{12,6 + \dots + 13,1}{8} = 13 \text{ libras-força}$$

- A média pode ser interpretada com uma medida de localização.
- Uma média $\bar{x} = 13$ pode ser pensada com ponto de balanço.
- Pode ser vista como centro de massa da amostra.

Figure 6-1 The sample mean as a balance point for a system of weights.



- A média da amostra é:
 - o valor médio de todas observações do conjunto de dados.
- Essa **amostra** foi selecionada a partir de uma **população**.
- No exemplo, a população são todos conectores que serão fabricados.
- É uma **população hipotética** $\Rightarrow$ não existe fisicamente.
- As vezes existe uma população física real:
 - as pastilhas de silicone produzidas em uma fábrica.

- Vimos no capítulo anterior a média μ de uma distribuição de probabilidade.
- Podemos pensar em uma distribuição de probabilidade como um modelo para uma população.
- A média é o valor médio de todas observações dessa população.
- Se a população tem um tamanho finito N de valores equiprováveis:
 - a função de probabilidade é

$$f(x_i) = \frac{1}{N}$$

- a média é

$$\mu = \sum_i x_i f(x_i) = \frac{\sum_{i=1}^N x_i}{N}.$$

- A média amostral nos dá uma estimativa da média populacional.
- No exemplo do conector:
 - o engenheiro concluiria que uma estimativa da força de remoção média é

13,0 libras- força .

- Essa informação foi obtida com base na amostra.

- Uma outra medida de tendência central é a **mediana**.
- Em um conjunto de dados ordenados:
 - a mediana separa a metade inferior e superior dos dados.
- Metade da população terá valores menores ou iguais a mediana.
- Metade da população terá valores maiores ou iguais a ela.

Como encontrar a mediana

- Ordene o conjunto de dados.
- Seja n o número total de observações.
- A mediana é dada por

$$M = \begin{cases} X_{((n+1)/2)} & \text{se } n \text{ é ímpar} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{se } n \text{ é par.} \end{cases}$$

Exemplo:

- Considere as seguintes observações

$$x_1 = 1 \quad x_2 = 3 \quad x_3 = 5 \quad x_4 = 7 \quad x_5 = 9 .$$

- Os dados já estão ordenados.
- Temos que $n = 5$ e a mediana é dada por

$$M = x_{(3)} = 5 .$$

Exemplo:

- Considere agora as seguintes observações

$$x_1 = 1 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 8 \quad x_5 = 9 \quad x_6 = 10 .$$

- Os dados já estão ordenados.
- Temos que $n = 6$ e a mediana é dada por

$$M = \frac{x_{(3)} + x_{(4)}}{2} = \frac{4 + 8}{2} = 6 .$$

Observações

- Em algumas situações a **mediana** deve ser usada no lugar da **média**.
- É menos afetada por *outliers* do que a média.
- Deve ser usada quando existem muitas observações discrepantes.

Exemplo:

- Queremos ter informação sobre o salário de uma determinada carreira.
- Uma média alta pode dar a falsa impressão de que os profissionais da classe tem um salário alto.
- Isso pode estar encobrindo o fato de que poucas pessoas possuem salários extremamente altos.
- Nesse caso, a mediana seria mais informativa.

- Podemos dividir o conjunto de dados em mais de duas partes.
- Depois de ordenar os dados podemos dividir em quatro partes:
 - obtemos assim os quartis.
- **Primeiro quartil (q_1):**
 - aproximadamente 25% das observações estão abaixo dele;
 - aproximadamente 75% das observações estão acima dele.
- **Segundo quartil (q_2):**
 - é a mediana.
- **Terceiro quartil (q_3):**
 - aproximadamente 75% das observações estão abaixo dele;
 - aproximadamente 25% das observações estão acima dele.

- Em geral o **percentil** $100k$ é tal que:
 - aproximadamente $100k\%$ das observações estão abaixo dele;
 - aproximadamente $100(1 - k)\%$ das observações estão acima dele.

Como encontrar quartis e percentis?

- Existem várias formas.
- Vamos apresentar uma mais simples.
- Dados: 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7.
- Temos que $n = 10$.
- Posição do $q_1 \Rightarrow 0,25 \times 11 = 2,75$

$$q_1 = \frac{2 + 2,1}{2} = 2,05.$$

- Posição do $q_2 \Rightarrow 0,5 \times 11 = 5,5$

$$q_2 = \frac{3 + 3,1}{2} = 3,05.$$

- Posição do $q_3 \Rightarrow 0,75 \times 11 = 8,25$

$$q_3 = \frac{3,7 + 6,1}{2} = 4,9.$$

- Uma outra medida de tendência central é o **ponto médio**

$$M = \frac{X_{(n)} + X_{(1)}}{2} .$$

- Não é muito utilizada.
- Essa medida perde muita informação dos dados.

- Outra medida de posição é a **moda amostral**.
- É o valor que ocorre com mais frequência.
- Considere os dados

1, 3, 5, 5, 6

a moda é 5.

- Alguns conjuntos de dado podem ter mais de uma moda.
- Pode ocorrer também de não haver nenhuma moda:

1, 3, 2, 5, 8, 7, 9.

Medidas de Variabilidade

- A média é muito útil
 - porém ela não fornece toda informação sobre os dados.
- Não basta sabermos a localização.
- Precisamos de ter uma ideia da variabilidade.
- Uma forma de medir a variabilidade da amostra:
 - **variância** amostral;
 - **desvio-padrão** amostral.

Variância da Amostra

- Uma amostra de tamanho n é coletada.
- As observações dessa amostra são denotada por

$$x_1, x_2, \dots, x_n.$$

- A **variância amostral** é dada por

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

- O **desvio-padrão** amostral é a raiz da variância.

- A unidade de medida da variância são quadrado das unidades originais.
- Se x é medida em libras, sua variância é medida em libras².
- O desvio-padrão possui a mesma unidade da variável x .

Como a variância da amostra mede variabilidade?

- A figura mostra os desvios $(x_i - \bar{x})$ para os dados da força de remoção do conector.
- Quanto maior a variabilidade dos dados maior será o valor absoluto de alguns desvios $(x_i - \bar{x})$.

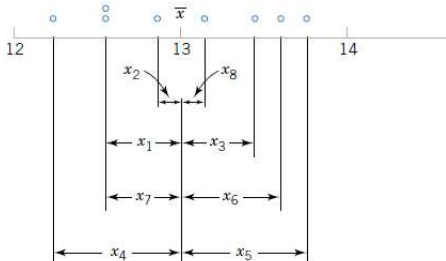


Figure 6-2 How the sample variance measures variability through the deviations $x_i - \bar{x}$.

- Não podemos apenas somar os desvios pois

$$\sum_i (x_i - \bar{x}) = 0 .$$

- Precisamos de uma medida que transforme os desvios negativos em positivos.
- Elevamos ao quadrado.
- Se a s^2 for pequena haverá relativamente pouca variabilidade nos dados.
- Se a s^2 for grande a variabilidade será relativamente alta.

Exemplo

- Considere o exemplo da força de remoção.
- A tabela abaixo apresenta as quantidades necessárias para o cálculo da variância.

Table 6-1 Calculation of Terms for the Sample Variance and Sample Standard Deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	-0.4	0.16
2	12.9	-0.1	0.01
3	13.4	0.4	0.16
4	12.3	-0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	-0.4	0.16
8	13.1	0.1	0.01
	<u>104.0</u>	<u>0.0</u>	<u>1.60</u>

Exemplo (continuação)

- Temos que

$$\sum_{i=1}^8 (x_i - \bar{x})^2 = 1,60$$

e a variância da amostra é

$$s^2 = \frac{1,60}{8-1} = \frac{1,60}{7} = 0,2286$$

e o desvio-padrão da amostra é

$$s = \sqrt{0,2286} = 0,48 \text{ libras-força.}$$

Outra forma de calcular s^2

- O cálculo de s^2 pode ser feito de uma maneira mais fácil.
- Basta observar que

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

- Ou seja

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

Exemplo:

- Vamos recalculer a variância do exercício anterior usando essa fórmula alternativa

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_i x_i)^2}{n}}{n - 1} = \frac{1353,6 - \frac{(104)^2}{8}}{7} = 0,2286.$$

- Vimos até agora como calcular a variância de uma **amostra**.
- Podemos estar interessados em calcular a variância da **população**.
- Vamos denotar a variância populacional por σ^2 .
- Quando a população tiver um tamanho N e todos valores são igualmente prováveis

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{N}$$

onde μ é a média da população.

- A variância da amostra pode ser usada como uma estimativa para variância populacional.

Observações:

- O divisor da variância amostral é o tamanho da amostra menos um ($n - 1$).
- O divisor da variância populacional é o tamanho da população N .
- Se o valor de μ fosse conhecido:
 - poderíamos calcular a variância amostral dividindo por n .
- Porém, como μ é desconhecido e estimado por $\bar{x}$
 - devemos dividir por $n - 1$.

- Muitas vezes é importante olharmos para variabilidade relativa.
- Ela depende da escala em que estamos medindo a variável.
- Usamos então o **coeficiente de variação**

Coeficiente de variação

- É o desvio padrão dividido pela média amostral.
- É expresso em porcentagem.

Exemplo:

- Medimos duas variáveis em uma amostra de alunos:
 - altura e peso.
- Os dados obtidos são

	Média	Desvio-padrão	Coeficiente de variação
Altura	1,143 m	0,063 m	5,5%
Peso	50 kg	6 kg	12%

- A variabilidade dos pesos dos alunos é aproximadamente o dobro da variabilidade de suas alturas.

Exemplo:

- Medimos a altura (cm) de dois grupos:
 - recém-nascidos e adolescentes.
- Obtemos o seguinte resultados

	Média	Desvio-padrão
Recém-nascidos	50	6
Adolescentes	160	16

- Qual grupo tem maior variabilidade?

Exemplo: (continuação)

	Média	Desvio-padrão	Coeficiente de variação
Recém-nascidos	50	6	12%
Adolescentes	160	16	10%

- A variabilidade do grupo de recém-nascidos é um pouco superior.

- Uma outra medida de variabilidade é a **amplitude**:
 - diferença entre a maior e menor observação.

Amplitude da Amostra

- Observamos uma amostra de tamanho n .
- As observações são

$$X_1, \dots, X_n .$$

- A amplitude da amostra é dada por

$$r = \max(x_i) - \min(x_i) .$$

Exemplo:

- Considere os dados da força de remoção.
- Nesse caso temos que

$$\min(x_i) = 12,3 \quad \max(x_i) = 13,6 .$$

- A amplitude então é dada por

$$r = \max(x_i) - \min(x_i) = 13,6 - 12,3 = 1,3 .$$

- À medida que a variabilidade aumenta
 - a amplitude da amostra aumenta.
- Porém ela ignora toda informação contida nos dados.
- Quando a amostra é muito pequena essa perda é ainda mais séria.
- A amplitude é muito usada em controle estatístico da qualidade.

- Uma outra medida de variabilidade é a **distância interquartílica**.
- É dada pela diferença entre o primeiro e terceiro quartil:

$$IQR = q_3 - q_1 .$$

- Para o exemplo mostrado antes

$$q_1 = 2,05 \quad \text{e} \quad q_3 = 4,9 .$$

$$IQR = 4,9 - 2,05 = 2,85 .$$

Observação

- Na maioria dos problemas em estatística trabalhamos com uma amostra.
- Essa amostra é coletada de uma população.
- Veja o esquema abaixo.

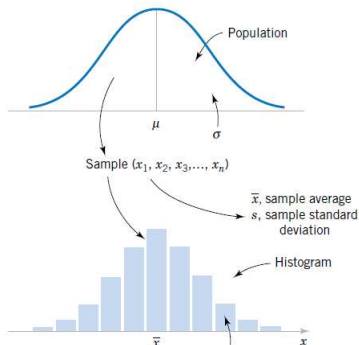


Figure 6-3 Relation-
ship between a popula-