

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TRANSPORTES
E GESTÃO TERRITORIAL – PPGTG
DEPARTAMENTO DE ENGENHARIA CIVIL – ECV

DISCIPLINA: TGT410026 – FUNDAMENTOS DE ESTATÍSTICA

3ª AULA: ESTATÍSTICA DESCRITIVA – Medidas Numéricas

Nos exemplos anteriores foi revisado como sintetizar os dados – Estatística Descritiva – utilizando os métodos tabulares e gráficos.

Na sequência revisaremos como calcular e interpretar algumas das estatísticas usadas. Além da média, da mediana e da moda revisaremos outras estatísticas descritivas, por exemplo, amplitude, desvio padrão, percentis e correlação. Tais estatísticas ajudarão na compreensão e interpretação dos dados.

1. MEDIDAS DE POSIÇÃO

Média da amostra: $\bar{x} = \frac{\sum x_i}{n}$ onde n é o tamanho da amostra e x_i são os valores observados (medidos)

Média da população: $\mu = \frac{\sum x_i}{N}$ onde N é o tamanho da população

Mediana: organizar os dados em ordem crescente; a) para um número ímpar de observações a mediana é o valor intermediário; b) para um número par de observações, a mediana é a média dos dois valores intermediários.

Moda: de um conjunto de dados é o valor que ocorre com maior frequência.

Percentis – quartis: aplicação restrita em engenharia ...

2. MEDIDAS DE VARIABILIDADE

Amplitude: por definição a amplitude é a diferença entre o maior valor e o menor valor de um conjunto de dados. Não diz muita coisa ...

Variância da amostra: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Variância da população: $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$

Desvio padrão da amostra: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \therefore s = \sqrt{s^2}$

Desvio padrão da população: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \therefore \sigma = \sqrt{\sigma^2}$

Coeficiente de variação: $cv = \left(\frac{s}{\bar{x}} \times 100 \right) \%$

Observação:

O coeficiente de variação de uma amostra nos informa qual o valor percentual do desvio padrão em relação a média da amostra. Em geral o cv é uma estatística útil para comparar a variabilidade de variáveis que possuem desvios padrão diferentes e médias diferentes.

Aplicações:

2.1 As pontuações de um jogador de boliche em seis jogos foram 182, 162, 184, 190, 170 e 174. Usando esses dados como uma amostra calcular:

- Amplitude: _____
- Variância: _____
- Desvio padrão: _____
- Coeficiente de variação: _____

Pontos - x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
162		
170		
174		
182		
184		
190		
Soma:		
Média: $\bar{x} =$		

2.2 Os preços de aluguel de carro por dia de uma amostra de sete cidades da região leste dos EUA são os seguintes:

Cidade	Taxa	Cidade	Taxa	Cidade	Taxa	Cidade	Taxa
Boston	U\$ 43	Atlanta	U\$ 35	Miami	U\$ 34	New York	U\$ 58
Orlando	U\$ 30	Pittsburg	U\$ 30	Washington	U\$ 36		

- Calcule a média: _____ .
a variância: _____ .
e o desvio padrão dos preços de aluguel de carros: _____ .
- Uma amostra similar de sete cidades da região oeste dos EUA apresentou um preço médio da amostra correspondente a U\$ 38 por dia de aluguel de carros. A variância e o desvio padrão foram 12,3 e 3,5, respectivamente. Discuta quaisquer diferenças entre os preços de aluguel de carros nas cidades do oeste e do leste dos EUA.

2.3 Foram registrados os seguintes tempos pelos corredores de 400 e 1.600 metros de uma equipe de atletismo de uma Universidade (tempo em minutos):

Tempo para 400m: 0,92 0,98 1,04 0,90 0,99 $s =$ _____ $cv =$ _____

Tempo para 1.600m: 4,52 4,35 4,60 4,70 4,50 $s =$ _____ $cv =$ _____

Depois de ver essa amostra de tempos de corrida, um dos treinadores comentou que os corredores de 400 metros apresentaram tempos mais constantes. Use o desvio padrão e o coeficiente de variação para sintetizar a variabilidade dos dados. O uso do coeficiente de variação indica que o comentário do treinador se justifica?

3. MEDIDAS DA FORMA DA DISTRIBUIÇÃO, DA POSIÇÃO RELATIVA E DETECÇÃO DE PONTOS FORA DA CURVA

FORMA DA DISTRIBUIÇÃO – uma medida numérica da forma de uma distribuição é a **assimetria**. A expressão analítica é dada por: $Ass = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$. Softwares como o Excel fornecem prontamente o valor da assimetria. Dados inclinados à esquerda, a assimetria é negativa; inclinados à direita, a assimetria é positiva; dados simétricos a assimetria é nula.

PADRONIZAÇÃO ou contagem-z: $z_i = \frac{x_i - \bar{x}}{s}$

TEOREMA DE CHEBYSHEV – pelo menos $(1 - 1/z^2)$ dos valores de dados devem estar contidos em z desvios padrão da média, em que z é qualquer valor maior que 1.

$z = 2 \rightarrow$ pelo menos 75% dos valores de dados estarão contidos nesse intervalo;

$z = 3 \rightarrow$ pelo menos 89% dos valores de dados estarão contidos nesse intervalo;

$z = 4 \rightarrow$ pelo menos 94% dos valores de dados estarão contidos nesse intervalo.

Exemplo: notas dos exames semestrais de 100 estudantes de um curso de estatística apresentam as seguintes informações: média 70 pontos e desvio padrão de 5 pontos.

- Quantos estudantes tiveram notas no exame entre 60 e 80 pontos? (Resp: 75%)
- Quantos estudantes tiveram notas entre 58 e 82 pontos? (Resp: 82,6%)

Regra empírica – **Distribuição normal**

A distribuição dos dados é uma normal quando apresenta uma forma similar a forma de morro ou sino. Para dados que possuem distribuição com essa forma pode-se dizer que:

- Aproximadamente 68% dos valores de dados estarão contidos num intervalo de um desvio padrão da média;
- Aproximadamente 95% dos valores de dados estarão contidos em um intervalo de dois desvios padrão da média;
- Aproximadamente 99,5% dos valores de dados estarão contidos em um intervalo de três desvios padrão da média.

Grandezas como tempo, volume, massa, velocidades, etc. apresentam distribuição de frequência normal.

Exemplo:

As embalagens de detergente líquido são preenchidas automaticamente em uma linha de produção. Os volumes de preenchimento possuem distribuição normal. Se a média dos volumes de preenchimento for 16 onças (473,12 ml) e o desvio padrão 0,25 onça (7,4 ml), pode-se tirar as seguintes conclusões: (* - 1 onça americana = 29,57 ml; 1 onça inglesa = 28,4 ml)

- aproximadamente 68% das embalagens cheias terão carga entre 15,75 e 16,25 onças (um desvio);
- aproximadamente 95% das embalagens cheias terão carga entre 15,50 e 16,50 onças (2 desvios);
- aproximadamente 99,5% das embalagens cheias terão carga entre 15,25 e 16,75 onças (3 desvios).

DETECÇÃO DE **PONTOS FORA DA CURVA**

Pontos fora da curva são relativos aos dados com valores excepcionalmente grandes ou pequenos. Antes de eliminá-los, devem-se analisar três hipóteses:

- Um ponto fora da curva pode ser um valor de dados que foi incorretamente registrado: neste caso é possível corrigir e prosseguir a análise.
- Um ponto fora da curva também pode ser proveniente de uma observação que foi

- incorretamente incluída no conjunto de dados: nesse caso ela deve ser **eliminada**.
- c) Pode ainda ser um valor de dados incomum que foi registrado corretamente e que pertence ao conjunto de dados. Nesses casos ele deve permanecer.
 - d) A padronização ou contagem-z permite analisar, com boa segurança, a existência de pontos fora da curva; essa técnica deve ser usada em conjunto com a regra empírica ou distribuição normal.

4. MEDIDAS DE ASSOCIAÇÃO ENTRE DUAS VARIÁVEIS

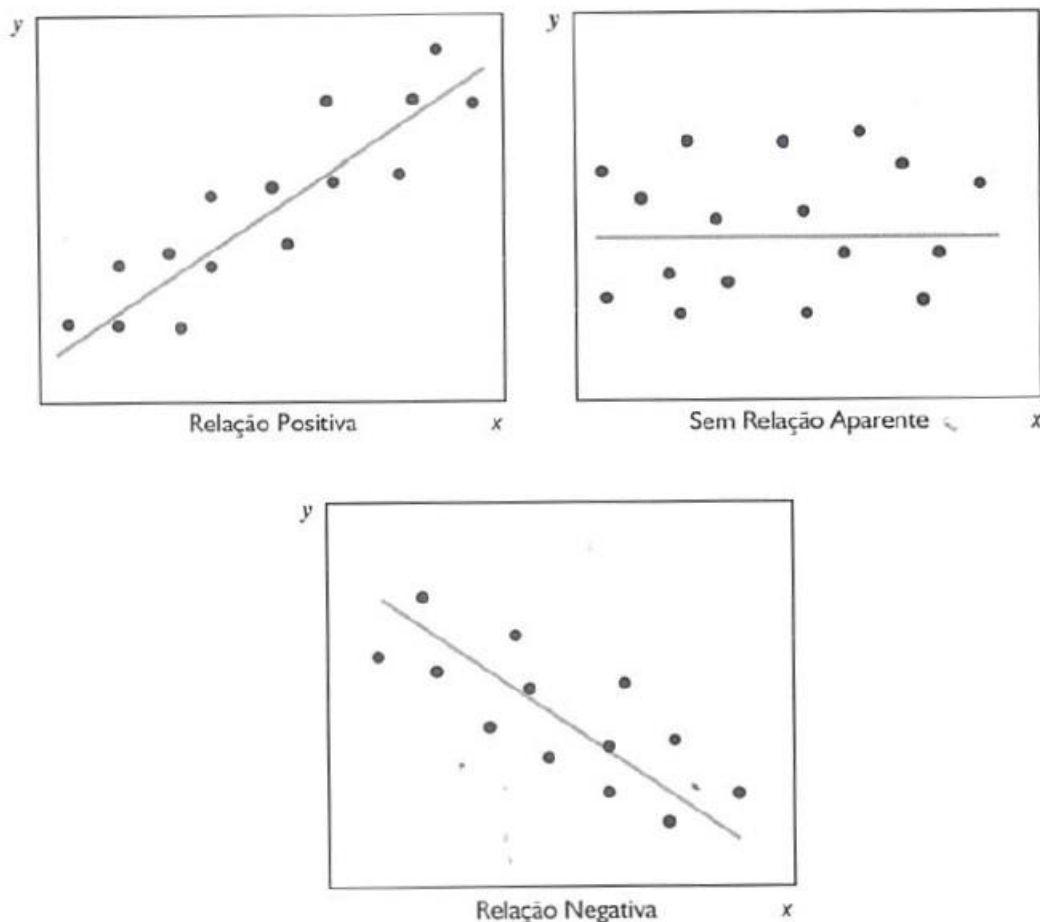
Agora, o interesse recai em conhecer o comportamento não de uma variável apenas, mas de duas variáveis. Em geral se quer saber qual a relação entre duas variáveis ou mais. As medidas descritivas da relação entre duas variáveis são: a *covariância* e a *correlação*.

Covariância – medida descritiva da associação linear entre duas variáveis; x_i e y_i são as observações das variáveis x e y . O tamanho da amostra em análise é designado por n .

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Interpretação da covariância:



Coeficiente de correlação – é um número que indica a relação linear entre duas variáveis. Usa o conceito de covariância entre as variáveis.

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Interpretação do coeficiente de correlação: r igual a ± 1 indica relação linear perfeita entre as variáveis; $r = 0$ indica nenhuma relação linear entre as variáveis. Para valores de r diferentes de zero e um, há indicação de alguma relação linear. Como exemplo verifique a relação entre as variáveis baseando-se nos seguintes dados amostrais:

xi	yi	$s_{xy} = 100$	$s_x = 5$	$s_y = 20$	$r_{xy} = 1$
5	10				
10	30				
15	50				

5. MÉDIA PONDERADA – VALORES INDIVIDUAIS E AGRUPADOS

Em algumas situações a média aritmética não reflete com precisão o valor médio. Pode ocorrer superestimação ou subestimação. Nestes casos a média ponderada é sempre mais indicada.

O problema na adoção da média ponderada é a fixação dos pesos de cada observação. O peso deve sempre refletir o grau de importância de cada observação no cálculo da média. Como exemplo clássico de uso da média ponderada é o cálculo da média escolar quando os resultados são expressos através de escores A (4), B (3), C (2), D (1) e F (0). Os pesos são o número de horas-crédito conquistadas para cada grau. Outro exemplo: os dados a seguir referem-se a cinco compras de determinada matéria prima ao longo dos últimos 3 meses. O gerente quer saber o custo médio por quilo da matéria prima.

Compra	Custo por quilo (R\$)	Quantidade (Kg)
1	3,00	1.200
2	3,40	500
3	2,80	2.750
4	2,90	1.000
5	3,25	800

O resultado apontou um custo médio de R\$ 2,96 por quilo; quais os pesos utilizados?

Dados agrupados – não raras vezes os dados para análise estatística se apresentam na forma agrupada ou na forma de distribuição de frequência. Nestes casos também é possível estimar a média e a variância (desvio padrão) de dados agrupados.

$$\bar{x} = \frac{\sum f_i \cdot M_i}{n}$$

$$\mu = \frac{\sum f_i \cdot M_i}{N}$$

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n - 1}$$

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N}$$

Exemplo: os dados a seguir sintetizam a distribuição de frequência dos tempos necessários para conclusão dos trabalhos de auditoria de uma determinada empresa. Qual é o tempo médio para a

conclusão das auditorias relativo à amostra?

Tempo (dias)	Frequência f_i	Ponto médio M_i	$f_i \cdot M_i$	Desvio $(M_i - \bar{x})$	$(\text{Desvio})^2$ $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
10-14	4					
15-19	8					
20-24	5					
25-29	2					
30-34	1					
Soma:						

Média da amostra: $\bar{x} = 19$ dias

Variância: $s^2 = 30$ dias

Desvio padrão: $s = 5,48$ dias

OBSERVAÇÃO: quando calculamos a estatística descritiva de dados agrupados utilizamos os pontos médios da classe para aproximar os valores de dados da classe; isso fornece apenas uma aproximação da estatística descritiva se usássemos os dados originais diretamente. Portanto, SEMPRE QUE POSSÍVEL USE OS DADOS ORIGINAIS PARA CALCULAR A ESTATÍSTICA DESCRITIVA.!!!!

EXERCÍCIOS E APLICAÇÕES

1. A *grade point average* (GPA) dos estudantes universitários norte-americanos baseia-se no cálculo de uma média ponderada. Na maioria das escolas de ensino superior (*colleges*) são atribuídos valores aos graus conforme descrito acima. Depois de 60 horas-crédito de trabalho acadêmico, um estudante de uma universidade pública ganhou 9 horas-crédito para A, 15 horas-crédito para B, 33 horas-crédito para C e 3 horas-crédito para D.

a) calcule o GPA do estudante;

b) os estudantes da escola pública precisam manter um GPA de 2,5 para suas primeiras 60 horas-crédito de trabalho acadêmico a fim de serem admitidos na escola de Administração. Esse estudante será admitido?

GPA (xi)	Pesos (wi)	(wi . xi)
4 (A)	9	
3 (B)	15	
2 (C)	33	
1 (D)	3	
0 (F)	0	
Soma:		

$$a) \bar{x} = \frac{\sum w_i \cdot x_i}{\sum w_i} =$$

b) É admitido? _____

2. Um posto de gasolina registrou a seguinte distribuição de frequência para o número de galões de gasolina vendidos por carro em uma amostra de 680 carros.

Gasolina galões	Frequência f_i	Ponto médio M_i	$f_i \cdot M_i$	Desvio $(M_i - \bar{x})$	$(\text{Desvio})^2$ $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
0-4	74					
5-9	192					
10-14	280					
15-19	105					
20-24	23					
25-29	6					
Soma:						

Calcule a média (10,74), a variância (25,63) e o desvio padrão (5,06) desses dados. Se o posto de gasolina espera atender cerca de 120 carros em determinado dia, estime o número total de galões de gasolina que serão vendidos (1.288,80).

3. Cinco observações feitas de duas variáveis são apresentadas a seguir:

Xi → 6 11 15 21 27

Yi → 6 9 6 17 12

- Desenvolva um gráfico de dispersão desses dados;
- O que o diagrama de dispersão indica a respeito da relação entre x e y?
- Calcule e interprete a covariância da amostra. (Resp: 26,5)
- Calcule e interprete a correlação da amostra. (Resp: 0,693)

4. Um estudo do Departamento de Transportes sobre a velocidade ao volante e a milhagem de automóveis de tamanho médio resultou nos seguintes dados:

Vel. ao volante 30 50 40 55 30 25 60 25 50 55

Milhagem 28 25 25 23 30 32 21 35 26 25

Calcule e interprete o coeficiente de correlação da amostra. (Resp: -0,91)

5. Suponha que os dados tenham uma distribuição em forma de sino (normal) com uma média igual a 30 e desvio padrão igual a 5. Determinar a percentagem de dados que se encontram dentro de cada um dos intervalos:

- 20 a 40
- 15 a 45
- 25 a 35

6. Os elevados custos praticados no mercado imobiliário da Califórnia fizeram com que as famílias que não podem se dar ao luxo de comprar casas maiores considerem as construções de quintal como uma alternativa de expandir suas residências. Muitas utilizam as estruturas existentes em seus quintais como escritórios, estúdios artísticos e áreas de lazer, bem como para armazenamento adicional. O preço médio de uma construção de quintal personalizada, feita em madeira e coberta com telhas de amianto é US\$ 3.100. Suponha que o desvio padrão seja de US\$ 1.200.

- qual é a contagem-z de uma estrutura de quintal que custa US\$ 2.300? _____
- qual é a contagem-z de uma estrutura de quintal que custa US\$ 4.900? _____
- Interprete as contagens-z dos itens (a) e (b). Comente se algum deles seria considerado fora da curva. _____.
- Um artigo de jornal descreveu a combinação de uma edícula-escritório construída em Albany, Califórnia, por US\$ 13.000. Essa estrutura deveria ser considerada fora da curva? Explique.

7) Os dados a seguir foram usados para construir os histogramas do número de dias necessários para a empresa A e a empresa B emitirem os pedidos de compra:

Empresa A – prazos (dias) 11 10 9 10 11 11 10 11 10 10

Empresa B – prazos (dias) 8 10 13 7 10 11 10 7 15 12

Use a amplitude e o desvio padrão para sustentar a observação de que a empresa A apresenta os prazos de entrega mais coerentes e confiáveis.

Empresa (A) Amplitude: _____ Desvio padrão: _____

Empresa (B) Amplitude: _____ Desvio padrão: _____

Relatório: _____
