Distribuições Bidimensionais

Apontamentos sobre o diagrama de dispersão, a covariância e o coeficiente de correlação

Page

- Em alguns estudos estatísticos há a incidência sobre 2 caracteres da mesma população, análises bivariadas, com a intenção de os comparar e ver se há ou não algum tipo de relações entre eles, ou se pelo contrário, são independentes. Em situações normais a variação de um dos caracteres influência a variação do outro.
    - Exemplo:
        - idade (em meses) e a massa corporal (em kg), dos 0 aos 24 meses;
        - idade (em meses) e o perímetro cefálico (em cm), dos 0 aos 36 meses;
        - idade (em anos) e a estatura (em cm), dos 2 aos 20 anos;
        - O número de trabalhadores a executar uma obra e o tempo de execução;
        - O rendimento mensal do agregado familiar e os gastos em lazer.

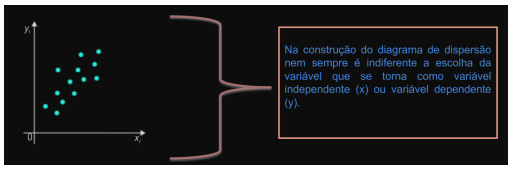
- No caso de se pretender 2 características conjuntamente, os dados observados aparecem sob a forma de pares de valores, isto é, cada indivíduo ou resultado experimental contribui com um conjunto de 2 valores.
- Deixamos de estar interessados em explorar isoladamente cada uma das variáveis
- A amostra de dados bivariados pode ser representada por:

$$(x_1, y_1), (x_2, y_2), ..., (x_i, y_i), ..., (x_n, y_n)$$

- A sua representação gráfica é feita através de um gráfico, designado por diagrama de dispersão ou nuvem de pontos

Diagrama de dispersão

- É uma representação gráfica para os dados bivariados, em que cada par de dados (xi, yi) é representada por um ponto de coordenadas (xi, yi), num sistema de eixos coordenados



Exemplo 1:

A tabela seguinte mostra os resultados obtidos por observação da temperatura em graus Celsius (ºC) e da pressão atmosférica em milímetros de mercúrio (mmHg), durante 7 dias

Temperatura (ºC)	18	20	21	19
17	21	22		
Pressão atmosférica (mmHg)	810	810	800	800
800	815	805		

1.1 Diga, justificando, se se trata ou não de dados bivariados

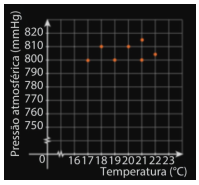
Resposta: São dados bivariados, pois a cada dia de observação corresponde um para de valores

1.2 Quais as variáveis em estudo?

Resposta: As variáveis em estudo são a temperatura, em ºC, e a pressão atmosférica, em mmHg

1.3 Representa a informação usando um diagrama de dispersão

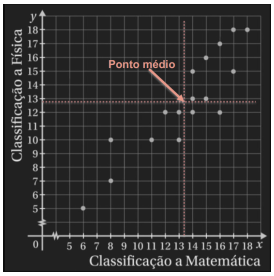
Resposta:

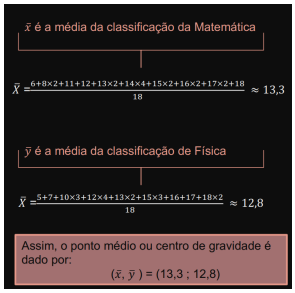


Centro de gravidade de uma distribuição bidimensional

- O centro de gravidade de uma distribuição bidimensional é o ponto G(x, y), onde x é a média dos valores da variável independente (xi) e y é a média dos valores da variável dependente (yi)

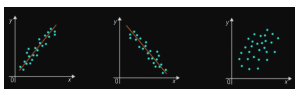
Exemplo 2:





Análise gráfica de dados

- Variáveis positivamente associadas
- Variáveis negativamente associadas
- Associação fraca entre variáveis



Covariância

- Para além dos indicadores numéricos que caracterizam individualmente cada uma das amostras (média, variância, desvio padrão, ...), podem-se definir novos parâmetros para descrever as relações existentes numa amostra bivariada
- Define-se covariância de x e y como:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Propriedades

- A covariância é um indicador da associação (linear) entre 2 variáveis:
  - quando cov(x, y) > 0, há correlação positiva
  - quando cov(x, y) < 0, há correlação negativa

- A covariância tem, no entanto, um forte inconveniente: depende da unidade de medida usada, sendo fortemente afetada por mudanças de escala nas observações

Coeficiente de correlação

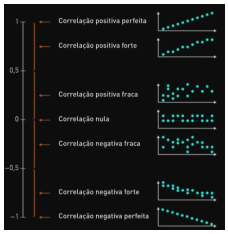
- Uma das medidas estatísticas que permite estabelecer o grau de correlação existente entre as variáveis é denominada coeficiente de correlação, representa-se por r e toma valores pertencentes ao intervalo [1, 1]
- Define-se coeficiente de correlação de uma amostra bivariada como:

$$r = r_{xy} = \frac{cov(x,y)}{S_x S_y}$$

$$r = \frac{n \times \sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{[n \times \sum x_i^2 - (\sum x_i)^2][n \times \sum y_i^2 - (\sum y_i)^2]}}$$

Interpretação do valor de r:

- Se r < 0, a correlação é negativa. A variação das variáveis é feita em sentidos opostos, isto é, uma aumenta quando a outra diminui
- Se r > 0, a correlação é positiva. A variação das variáveis é feita no mesmo sentido, isto é, uma aumenta quando a outra também aumenta.
- Se r = 0, a correlação é nula



- r = 1, se todos os pontos observados se encontram sobre uma reta de declive positivo
- r ≈ 1, se todos os pontos observados se encontram próximos de uma reta de declive positivo
- r ≈ 0, a nuvem apresenta um aspeto arredondado ou alongado segundo um dos eixos
- r ≈ -1, se todos os pontos observados se encontram próximos de uma reta de declive negativo
- r = -1, se todos os pontos observados se encontram sobre uma reta de declive negativo

- Um valor de r elevado não significa, necessariamente, uma associação linear forte
  - Pode ser uma consequência da estrutura da nuvem de pontos ou da existência de pontos afastados

- r ≈ 0 não significa mais do que a ausência de qualquer relação ou tendência linear entre as variáveis
  - Uma das variáveis pode ser completamente determinada pela outra e a correlação ser nula

- Não confundir associação estatística com causalidade:
  - Um valor elevado de r não significa que x seja causa de y ou que y seja causa de x

📈 Associação entre 2 variáveis