Metodologia de Diagnóstico e Elaboração de Relatório – FASHT

Medidas de dispersão e assimetria

Profª Cesaltina Pires
cpires@uevora.pt

Plano da Apresentação

- **Medidas de dispersão**
  - Variância
  - Desvio padrão
  - Erro absoluto médio
  - Coeficiente de dispersão
- **Índice de concentração e curva de Lorenz**
- **Medidas de assimetria**

Medidas de dispersão

- É muito diferente ter uma situação em que o salário médio mensal é 1000 euros e todos ganham 1000 euros, ou ter o mesmo salário médio mas em que metade das pessoas ganha 0 e outra metade ganha 2000 euros.
- Para caracterizar um conjunto de dados é importante não só a média, mas também a dispersão dos valores em torno da média (os valores estão todos próximos da média ou, pelo contrário, há valores muito afastados da média?)

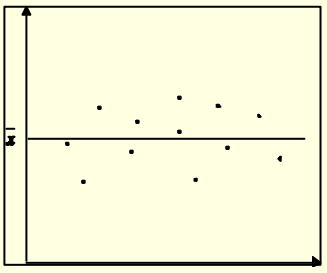
→ Qual o comportamento do conjunto de desvios em relação à média?

Como medir a dispersão?

Não podemos limitar-nos a somar os desvios em relação à média. Porquê?

→ Porque soma dos desvios é zero

Considerar medida que não leve em conta o sinal dos desvios (o que importa é a magnitude do desvio)



Considerar valor absoluto dos desvios → Desvio absoluto médio

Considerar o quadrado dos desvios → Variância

Desvio absoluto médio

O desvio médio é a média do valor absoluto dos desvios em relação à média

$$d = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Desvio médio

Aluno	Nota	Desvio	Desvio absoluto
1	12	-2	2
2	15	+1	1
3	11	-3	3
4	18	+4	4
5	17	+3	3
6	14	0	0
7	12	-2	2
8	13	-1	1
Soma	122	0	16

Desvio médio = 16/8 = 2

Variância

A variância é a média do quadrado dos desvios em relação à média:

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

Variância

Aluno	Nota	Desvio	Quadrado do desvio
1	12	-2	4
2	15	+1	1
3	11	-3	9
4	18	+4	16
5	17	+3	9
6	14	0	0
7	12	-2	4
8	13	-1	1
Soma	122	0	44

Variância = 44/8 = 5.5

Desvio padrão

O desvio padrão é a raiz quadrada da variância:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

Variância e desvio padrão em pequenas amostras

Quando a amostra é pequena deve calcular-se a variância e o desvio padrão corrigidos. A única diferença é que se divide por (*n*-1) em vez de dividir por *n*.
Por exemplo, a variância corrigida é:

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

Medidas de dispersão

- Tanto o desvio médio, como a variância e o desvio padrão são sensíveis à presença de outliers
- Outras medidas de dispersão:
  - Diferença entre extremos:

    valor máximo – valor mínimo

    É uma medida pouco robusta.
  - Dispersão quartal:

    3º quartil – 1º quartil

    Ou seja, é a amplitude do intervalo onde se situam as 50% observações centrais

Medidas de dispersão relativa

- As medidas de dispersão que analisamos até agora são todas sensíveis às unidades de medida. São medidas de dispersão absoluta.
- Pode ser útil ter medida que não dependa das unidades de medida (por exemplo: quando se querem comparar distribuições)

→ Medidas de dispersão relativa

Medidas de dispersão relativa

Coeficiente de dispersão

$$\frac{s}{\overline{x}}$$

Coeficiente de variação – é o coeficiente de dispersão × 100 (%)

Coeficiente de dispersão quartal – é a dispersão quartal / mediana

Índice de concentração

- Em muitos casos é interessante estudar o grau de concentração de uma dada variável.
- Exemplo: como é que a riqueza de um país está distribuída pelos seus cidadãos. Será que está igualmente distribuída, ou será que uma fracção grande dessa riqueza está concentrada em poucos indivíduos?
- Exemplo: no total dos pontos obtidos pelos clubes da super liga, como é que esses pontos são distribuídos? Será que há um pequeno grupo de clubes que tem a maioria dos pontos?
- Exemplo: no nº total de dias por ano que os trabalhadores faltam ao emprego, como é que esses dias se distribuem pelos vários trabalhadores. Será que todos os trabalhadores faltam um igual nº de dias, ou será que há um pequeno grupo de trabalhadores que são responsáveis por uma grande fracção no nº total de dias de absentismo?

Como medir a concentração?

Absentismo (dias)	Frequência absoluta	Absentismo total (por classe)	Frequência Relativa acumulada	Absentismo relativo acumulado
[0,2]	15383	10184	0.53	0.12
[3,5]	8592	31921	0.82	0.51
[6,10]	4546	31843	0.98	0.89
[11,20]	567	8762	1	1
Soma	**29088**	**82710**		

➡ 53% dos trabalhadores tem só 12% do total das faltas ao emprego, …

15

Como medir a concentração?

- Se não houvesse concentração, 10% das observações deviam ter 10% do total da variável em análise, 20% das observações deviam corresponder a 20% do total da variável, …
- Ideia é comparar andamento das frequências relativas acumuladas com andamento do valor relativo acumulado da variável em análise.
- Se representarmos graficamente o valor relativo acumulado da variável (eixo dos y's) em relação à frequência relativa acumulada (eixo dos x's) obtemos a curva de Lorenz.

Índice de Gini

Consideremos uma distribuição de frequências com k classes. Seja t_j o total do atributo correspondente aos n_j elementos da classe j. Se definirmos:

$$p_i = \frac{\sum_{j=1}^{i} n_j}{n} \quad \text{e} \quad q_i = \frac{\sum_{j=1}^{i} t_j}{t}$$

O *índice de concentração de Gini* é dado por:

$$G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

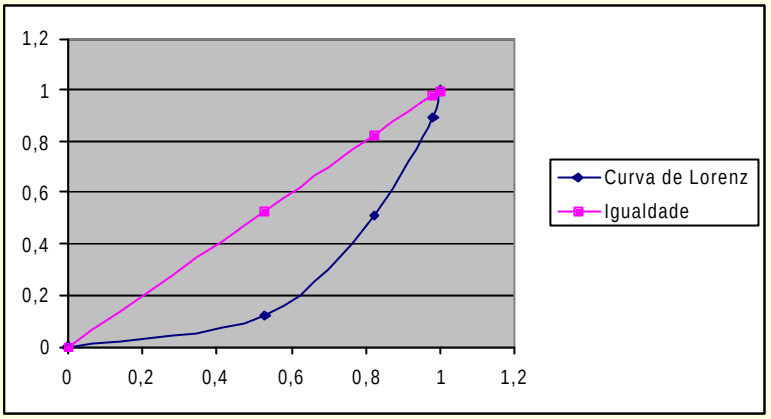
G toma valores entre 0 e 1 e é crescente com a concentração
G = 0 se houver igual repartição
G = 1 quando há concentração máxima

Índice de Gini

Absentismo (dias)	Frequência absoluta	Absentis. Total (por classe)	Frequência Relativa acumulada	Absentismo relativo acumulado
[0,2]	15383	10184	0.53	0.12
[3,5]	8592	31921	0.82	0.51
[6,10]	4546	31843	0.98	0.89
[11,20]	567	8762	1	1
Soma	**29088**	**82710**		

$$G = \frac{(0.53 - 0.12) + (0.82 - 0.51) + (0.98 - 0.89)}{0.53 + 0.82 + 0.98}$$

Curva de Lorenz



Medidas de Assimetria

- Como é que os valores se distribuem em relação ao «centro», de forma simétrica ou não?
- A observação do polígono de frequências ajuda a ver se distribuição é simétrica ou não.
- Nas distribuições simétricas média, mediana e moda coincidem. Nas distribuições assimétricas a média é «puxada» para o lado mais longo da distribuição
  - Assimetria positiva: média > mediana > moda
  - Assimetria negativa: média < mediana < moda

Medidas de Assimetria

Coeficiente de assimetria de Pearson:

$$g = \frac{\overline{x} - \text{moda}}{s}$$

Coeficiente de assimetria de Bowley:

$$g' = \frac{(F_{75\%} - F_{50\%}) - (F_{50\%} - F_{25\%})}{(F_{75\%} - F_{50\%}) + (F_{50\%} - F_{25\%})}$$