PROGRAMA
e
Metas Curriculares
Matemática A

Estatística

António Bivar, Carlos Grosso, Filipe Oliveira , Luísa Loura e
Maria Clementina Timóteo

GOVERNO DE PORTUGAL
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
Metas Curriculares

O tema da Estatística nos Cursos Científico-Humanísticos de Ciências e Tecnologias e de Ciências Socioeconómicas

Na interpretação do mundo real a Estatística procura responder a dois objetivos:

•Descrever de forma sucinta, por recurso a estatísticas resumo ou representações gráficas, informação censitária recolhida sobre uma certa "população" – **Estatística Descritiva**;

•Inferir para a população os padrões e indicadores estatísticos identificados na informação recolhida sobre uma sua parte (amostra) associando, em simultâneo, uma medida probabilística do erro que se poderá estar a cometer ao fazer essas inferências – **Estatística Inferencial**.

Ao longo do ensino básico, o tema da estatística foi sempre trabalhado no âmbito da Estatística Descritiva e, nesse sentido, terá transmitido aos alunos, não só as técnicas básicas de organização e tratamento de dados como, também, algum do seu cunho "criativo e artístico".

Optou-se, por isso, neste novo programa de Matemática A, por aprofundar, do ponto de vista das suas propriedades matemáticas, as estatísticas resumo basilares para a Estatística Inferencial: média, variância e quantis.

Metas Curriculares

Média e Variância

Terminologia e notação

Sejam:

- x uma variável estatística quantitativa em determinada população
- A uma amostra de dimensão $n \in \mathbb{N}$ dessa população

Admita-se que os elementos de A estão numerados de 1 a n

Represente-se por «x_i» o valor da variável x no elemento de A com o número i

$\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ designa-se por **amostra** da variável estatística x (ou, simplesmente amostra)

Metas Curriculares

Média e Variância

Terminologia e notação

Média

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Desvio de x_i em relação à média

$$d_i = x_i - \bar{x}$$

Soma dos quadrados dos desvios

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

Variância

$$s_x^2 = \frac{SS_x}{n-1}$$

Propriedades

$$\underset{\sim}{y} = (ax_1 + h, ax_2 + h, \ldots, ax_n + h)$$

$$\Downarrow$$

$$\bar{y} = a\bar{x} + h$$

e

$$s_y^2 = a^2 s_x^2$$

$$SS_x = 0 \text{ sse } x_1 = x_2 = \cdots = x_n$$

$$\therefore$$

$$s_x^2 = 0 \text{ sse } x_1 = x_2 = \cdots = x_n$$

$$SS_x = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Metas Curriculares

Média e Desvio Padrão como medidas de localização e dispersão

A importância do par (média, desvio padrão) em estatística, em deterimento de outras medidas de localização e de dispersão, como a mediana e a amplitude interquartis, por exemplo, decorre principalmente da relativa facilidade com que se demonstram algumas propriedades úteis.

De entre estas, considerou-se pertinente integrar no programa a conhecida desigualdade de Chebycheff, aqui enunciada no contexto de amostras de uma variável estatística.

Dada uma amostra $(x_1, x_2, \ldots, x_n)$ de desvio padrão não nulo, para qualquer *k* positivo, a percentagem de unidades estatísticas com valores **fora do intervalo**

$$]\bar{x} - ks_x \; ; \; \bar{x} + ks_x[$$

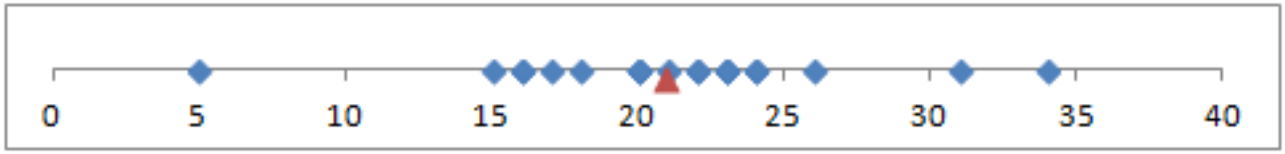
é sempre **menor ou igual** a $1/k^2$.

Metas Curriculares

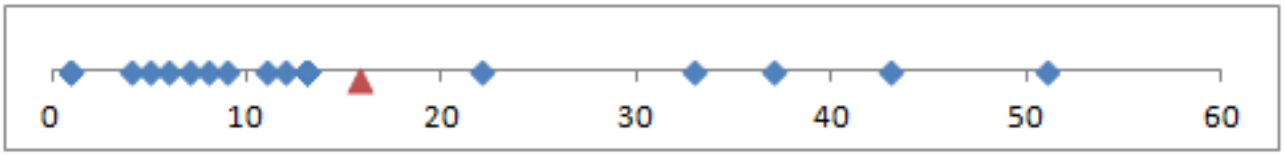
Média e Desvio Padrão - medidas pouco "resistentes"

As boas propriedades calculatórias da média e do desvio padrão foram determinantes no seu uso generalizado em estatística embora se lhes reconhecesse "debilidades" enquanto medidas resumo da localização e da dispersão da amostra.

Exemplo 1: distribuição não enviesada com uma zona central de densidade mais elevada. A média é um bom identificador dessa zona central de maior densidade .



Exemplo 2: distribuição enviesada com uma zona de densidade mais elevada junto dos pequenos valores. A média não é um bom identificador da zona de maior densidade.

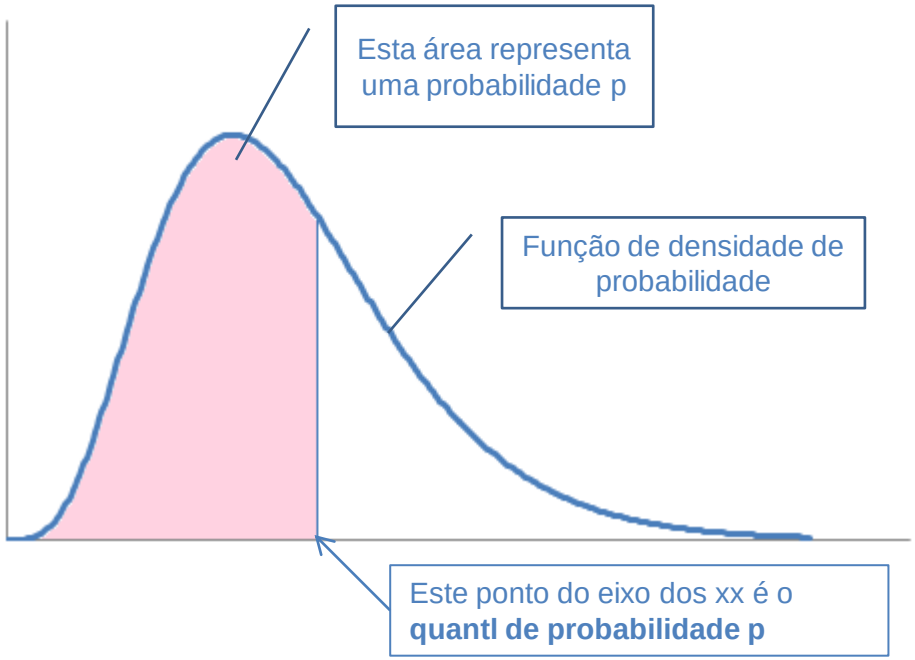


Nota: ver desenvolvimento sobre este tópico no caderno de apoio do 10.º ano.

Metas Curriculares

Percentil de ordem k como caso particular de um quantil

Definição de quantil (populacional) de probabilidade p



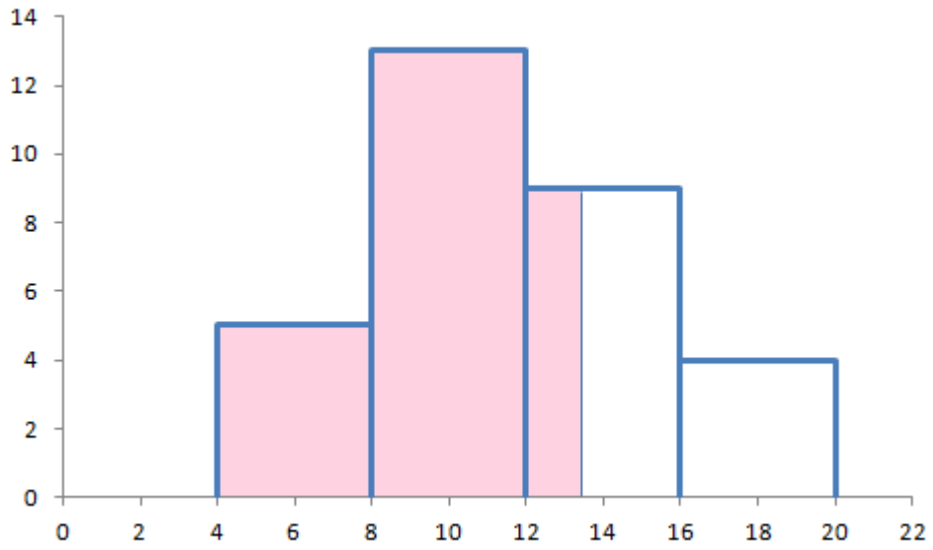
Em modelos de probabilidade que admitem função de densidade, a definição de quantil p é simples: toma-se a função que a cada ponto do eixo dos xx faz corresponder a área assinalada a rosa (função de distribuição); **o quantil de probabilidade p é o valor em p da inversa da função de distribuição**.

Em modelos de probabilidade que não admitem função de densidade, a definição de quantil p é análoga mas recorre à inversa generalizada da função de distribuição uma vez que, nesse caso, esta não é injetiva.

Metas Curriculares

Percentil (amostral) de ordem k

Ao nível do 10.º ano de escolaridade, a noção de percentil poderá ser introduzida, não através da função de densidade mas, sim, usando um exemplo de dados organizados na forma de **histograma**.



O percentil de ordem 65 ou, simplesmente, percentil 65, por exemplo, é o ponto do eixo horizontal para o qual a área acumulada dos retângulos do histograma que estão à sua esquerda, acrescida da área do retângulo que o ponto determina na classe a que pertence, é igual a 80% da área total do histograma.

Observação: quando se diz "percentil de ordem k", o "k" é um número natural menor ou igual a 100.

Metas Curriculares

Percentil (amostral) de ordem k: definição

Fazendo de novo o paralelismo entre a definição de percentil amostral e a definição de percentil populacional e relembrando que foi necessário fazer uma generalização ao passar de modelos com função de densidade de probabilidade para modelos em que esta não está definida, também no caso amostral se torna necessário adaptar a definição aos casos em que não faça sentido organizar os dados quantitativos na forma de histograma (note-se que o histograma é uma representação gráfica que só é adequada a dados de natureza contínua, muitas vezes designados também por "dados de medição").

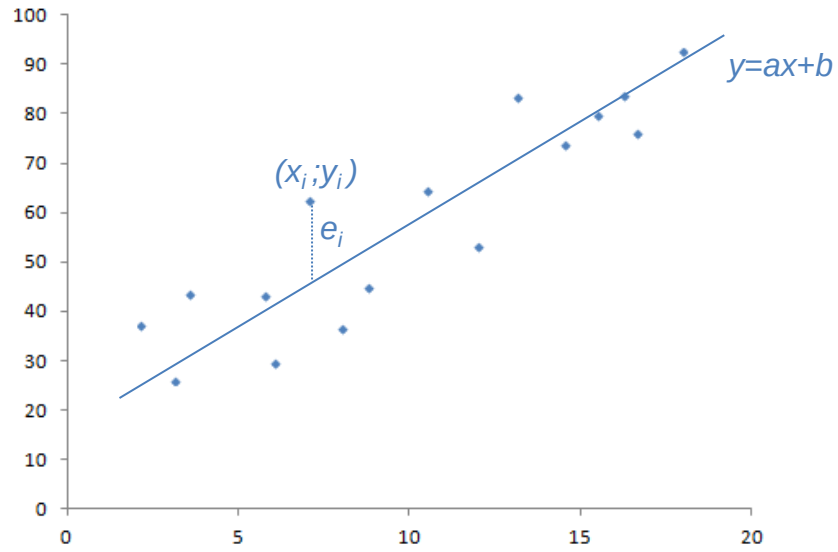
A adaptação da definição de percentil amostral ao caso geral não é única e, no âmbito do programa de Matemática A, optou-se por aquela que faz coincidir o percentil 50 com a mediana, tal como esta é definida ao nível do ensino básico.

O percentil de ordem k da amostra $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ define-se como:

- O valor máximo da amostra se $k = 100$
- A média dos elementos de ordem $\frac{kn}{100}$ e $\frac{kn}{100} + 1$ na amostra ordenada se $k \neq 100$ e $\frac{kn}{100}$ for inteiro
- O elemento de ordem $\left[\frac{kn}{100}\right] + 1$ na amostra ordenada, nos restantes casos

Metas Curriculares

Reta de mínimos quadrados



Desvio vertical do ponto $P_i(x_i, y_i)$ em relação à reta

$e_i = y_i - (ax_i + b)$

A reta de mínimos quadrados é aquela para a qual é mínima a soma dos quadrados dos desvios verticais. Dito de outro modo, é aquela cujo declive "a" e ordenada na origem "b" são tais que a função

$$f(a,b) = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

toma o valor mínimo. Uma vez que o estudo de extremos de funções de duas variáveis não faz parte do programa de secundário, optou-se por restringir a pesquisa ao conjunto das retas com ordenada na origem da forma $b = \bar{y} - a\bar{x}$, a que corresponde uma soma dos desvios verticais igual a zero.

A função acima passa a depender apenas de uma variável e a determinação do seu mínimo não mais será do que um exercício com a particularidade de envolver o símbolo de somatório.

Metas Curriculares

Dados estatísticos bivariados: descritores 1.4 a 1.7

4. Identificar, dadas duas variáveis estatísticas quantitativas x e y em determinada população e uma amostra A de dimensão $n \in \mathbb{N}$ dessa população cujos elementos estão numerados de 1 a n, a «amostra bivariada das variáveis estatísticas x e y» (ou simplesmente «amostra de dados bivariados (quantitativos)») como a sequência $((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n))$, representá-la por «$\underset{\sim}{(x,y)}$» e designar por «dimensão da amostra bivariada» o número natural n.
5. Determinar, em casos concretos de amostras de dados bivariados, qual das variáveis estatísticas deverá ser tomada como independente e qual deve ser tomada como dependente, utilizando argumentos que envolvam o conhecimento empírico das condicionantes físicas (ou outras) que poderão ter determinado a estrutura de relação entre as duas variáveis estatísticas.
6. Designar, dada uma amostra de dados bivariados, a variável considerada dependente por «variável resposta» e a variável considerada independente por «variável explicativa».
7. Designar, fixado um referencial ortonormado num plano, $n \in \mathbb{N}$ e uma amostra de dados bivariados quantitativos $\underset{\sim}{(x,y)} = ((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n))$, por «nuvem de pontos» o conjunto $\{(P_1(x_1,y_1),P_2(x_2,y_2),\dots,P_n(x_n,y_n)\}$ e saber que uma análise visual e intuitiva da nuvem de pontos poderá permitir argumentar se será ou não adequada a interpretação da relação entre as duas variáveis estatísticas através do ajustamento da reta de mínimos quadrados.

Metas Curriculares

Coeficiente de correlação linear

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{SS_x SS_y}}$$

Uma vez que o declive da reta de mínimos quadrados é dado por:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{SS_x}$$

facilmente se estabelece a relação entre o coeficiente de correlação e o declive da reta

$$r = a\sqrt{\frac{SS_x}{SS_y}}$$

O coeficiente de correlação tem, pois, sinal idêntico ao do declive da reta de mínimos quadrados.

Metas Curriculares