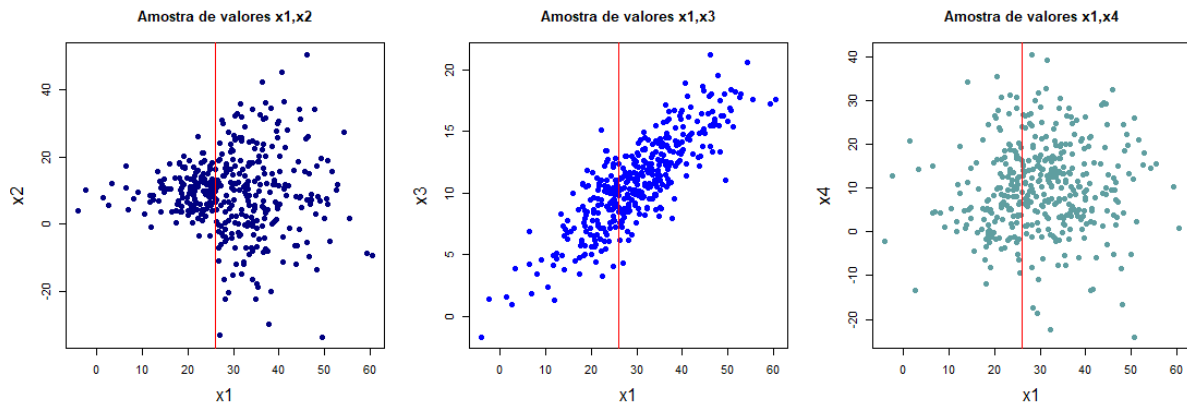




Workshop Exploração gráfica de dados de expressão genética usando R: uma introdução

XVIII ENEEB - Encontro Nacional de Estudantes de Engenharia Biomédica Universidade de Aveiro 23-26/fevereiro/2023

1. Considere uma amostra constituída por 400 observações relativas a quatro variáveis (x_1, x_2, x_3, x_4). Os dados da amostra relativa ao par de variáveis (x_1, x_i), com $i = 2, 3, 4$ foram representados usando diagramas de dispersão



A amostra total (400 observações) foi dividida em duas partes:

Subamostra 1: observações com $x_1 \leq 26$ (observações à esquerda da linha vertical)

Subamostra 2: restantes (observações à direita da linha vertical).

- (a) Entre as variáveis x_2, x_3, x_4 , identifique aquela(s) que tende(m) a apresentar a seguinte propriedade:
- a média mantém-se igual nas suas subamostras;
 - a média é (substancialmente) superior na subamostra 2;
 - a variabilidade mantém-se igual nas suas subamostras;
 - a variabilidade é substancialmente superior na subamostra 2.
- (b) Confronte a resposta dada na alínea anterior com a seguinte tabela de estatísticas sumárias:

Amostra	Variável		
	x_2	x_3	x_4
Total	$\bar{x} = 9.8$ $s_c = 12.4$ $amp = 94.2$	$\bar{x} = 11.0$ $s_c = 3.8$ $amp = 22.3$	$\bar{x} = 10.1$ $s_c = 10.1$ $amp = 64.4$
Subamostra 1	$\bar{x} = 9.5$ $s_c = 4.8$ $amp = 24.5$	$\bar{x} = 7.8$ $s_c = 2.8$ $amp = 15.5$	$\bar{x} = 9.0$ $s_c = 9.6$ $amp = 48.9$
Subamostra 2	$\bar{x} = 9.9$ $s_c = 15.3$ $amp = 94.2$	$\bar{x} = 13.0$ $s_c = 2.9$ $amp = 15.3$	$\bar{x} = 10.7$ $s_c = 10.4$ $amp = 64.4$

Nota: amp = amplitude da amostra

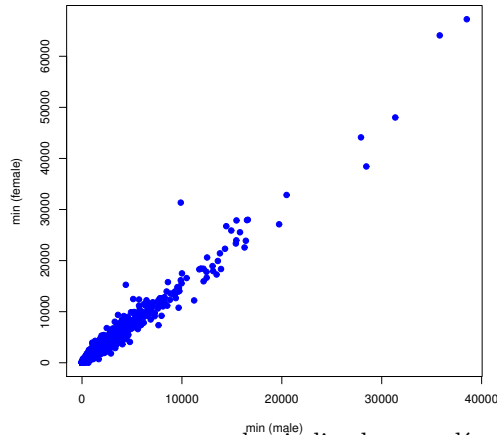
2. Faça a importação dos dados do ficheiro **Aveiro.txt**¹ para o **R**. Trata-se de um ficheiro contendo a contagem de 20738 marcadores genéticos (variáveis) de 6 pacientes (observações), 4 do sexo masculino (**m1,m2,m3,m4**) e 2 do sexo feminino (**f1,f2**), estando cada marcador identificado pela variável **ENTREZID**.
Observe que no ficheiro **Aveiro.txt** em linha estão as variáveis e em coluna estão as observações. Esta situação pode ocorrer quando se manipulam dados de elevada dimensionalidade com tamanho reduzido de amostras ($p \gg n$).
Nestas circunstâncias, a base de dados a usar no **R** deverá ser transpostas previamente! No **R** designe essa base de dados (indivíduos \times variáveis) por **dados.ficheiro**.

¹Dados cordialmente cedidos em 2019 pelo centro de investigação iBiMED da Universidade de Aveiro.

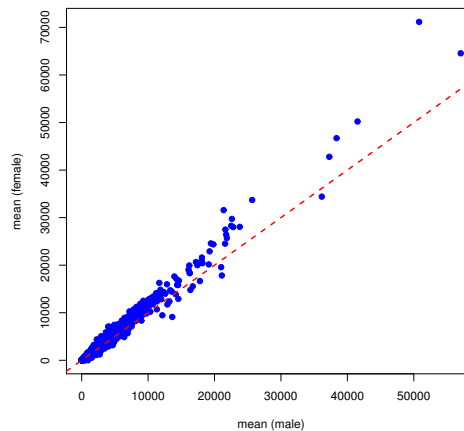
- (a)  Interprete o resultado da seguinte sequência de comandos:


```
> dadosA= data.frame(t(dados_ficheiro))
> Min.m=sapply (dadosA[2:5,], min)
> Min.f=sapply (dadosA[c(6,7),], min)
> plot(Min.m,Min.f,pch=19,col="red",xlab="min(male)",
+ ylab="min(female)")
```


- (b) Depois de ter executado mais alguns comandos, o investigador apresentou o gráfico seguinte. Compare-o com o gráfico obtido na alínea anterior e refira o que o investigador terá feito. Comente a estratégia usada.



- (c) Um outro investigador executou os mesmos comandos indicados na alínea 1. mas, em vez de considerar a função `min`, usou a função `mean` e eliminou as duas primeiras variáveis do ficheiro de dados. Obteve o seguinte gráfico (reta $y = x$ a tracejado). Assinale uma possível conjectura para este conjunto de dados.



Considere os dados `leukemia` do pacote `plsgenomics` do . Trata-se de uma base de dados contendo os níveis de expressão genética de 3051 genes (variáveis) de 38 amostras de mRNA de tumores (indivíduos) obtidos de um estudo usando microarrays realizado por Golub et al.(1999)². Existem dois tipos de tumor: *acute myeloid leukemia* (AML) e, em maior número, *acute lymphoblastic leukemia* (ALL).

- (a)  Familiarize-se com o dados e identifique quantos pacientes têm tumor do tipo AML. Para tal sugere-se que use os seguintes comandos:

```
> ?leukemia
> str(leukemia)
> table(leukemia$Y)
```

- (b)  Construa um `heatmap` simples para visualizar os dados fazendo:

```
> heatmap(leukemia$X,Rowv=NA, Colv=NA, main="leukemia data")
```

Consegue detetar a existência de padrão diferenciado entre os tipos de tumor?

²Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M. et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537