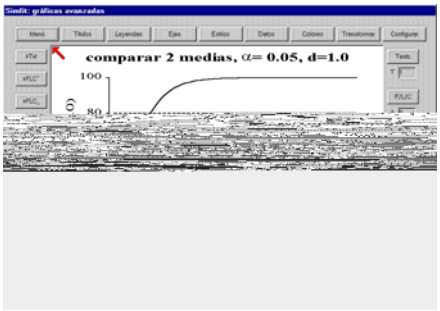
Importância da análise de dados em investigação
Possibilidades do pacote estatístico *SIMFIT* em Ciências da Vida

F. J. Burguillo[1] y W. G. Bardsley[2]

[1]*Fac. de Farmacia, Universidad de Salamanca, burgui@usal.es*
[2]*School of Biological Sciences, University of Manchester (U. K.), bill.bardsley@man.ac.uk*

Resumo: A análise de dados *on-line* tornou-se uma técnica indispensável em qualquer investigação em Ciências da Vida. No entanto, para usar bem as novas possibilidades da informática, não basta ter os programas apropriados, é necessário também saber os fundamentos estatísticos em que o programa é baseado. O objectivo deste relatório é comentar de um modo panorâmico estes fundamentos teóricos, a sua utilidade em qualquer investigação, e a sua projecção em Bioinformática. Para isso serão expostos, primeiro, as diferentes etapas que é necessário seguir para fazer uma boa análise de dados: desenho da experiência, exploração dos dados, análise do mesmo, e o uso de técnicas estatísticas especiais, se for necessário. Finalmente, serão expostos alguns exemplos de como o pacote estatístico SIMFIT nos pode ajudar em todas estas tarefas. SIMFIT foi desenvolvido por um de nós (W.G. Bardsley), é completamente compatível com Windows 98/2000/XP e está disponível em espanhol no endereço http://simfit.usal.es e em inglês no endereço http://www.simfit.man.ac.uk.

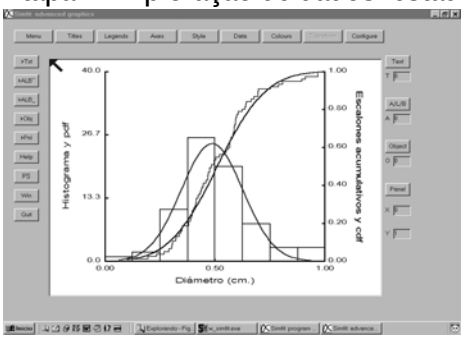
Etapa 1. Desenho de experiências: tamanho das amostras, n.º de pontos, intervalo.



Quando se quer desenhar experiências em que se vai aplicar um teste estatístico, por exemplo do tipo comparação de 2 médias mediante o teste "t", terá que se calcular "a priori" o *tamanho da amostra* necessária. Este tamanho, será o mínimo número de casos que terá que se utilizar no estudo se se quer detectar uma determinada diferença entre as médias (d), se é que existe. Para isso, há que fixar o risco de rejem

de uma enzima, há que projectar a experiência tendo em conta a *margem* da concentração do substrato, o *n.º de pontos*, *o intervalo* entre eles, etc. A *simulação on-line* de todos estes factores pode ajudar muito a decidir entre as diferentes opções.

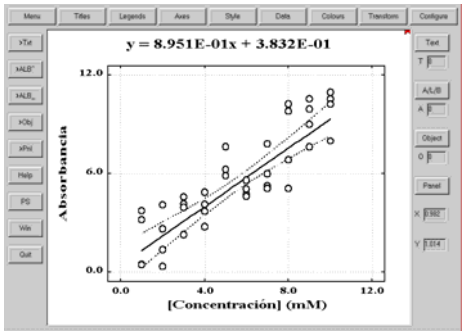
Etapa 2. Exploração de dados: estatística descritiva.



Em qualquer laboratório, é frequente a realização de medidas que flutuam ligeiramente devido ao azar, pelo que uma tarefa habitual é encontrar o valor médio, o desvio padrão, representar os dados em forma de histograma ou de curva cumulativa por passos, ajustá-los à função de distribuição normal....etc. *SIMFIT* pode fazer todo este tipo de cálculos e representações (figura da esquerda). Outras vezes, tem que se explorar a possível correlação entre variáveis (coeficientes de correlação de Pearson).

Etapa 3. Análises de dados: testes estatísticos, ajuste de curvas.

Testes estatísticos: SIMFIT realiza todas as provas estatísticas habituais, como o teste X-Quadrado sobre tabelas de contingência, teste de Fisher exacto em tabelas de 2 por 2, teste t de comparação de médias (para variâncias iguais ou distintas), ANOVA de 1, 2 ou 3 vias, teste Q de Tukey... etc. Dispõe também dos testes não paramétricos usuais, como a prova U de Mann-Whitney, o teste de Kolmogorov-Smirnov, o de Kruskal-Wallis, etc.



Ajuste de curvas de calibração e predição inversa: Existem em *SIMFIT* diferentes possibilidades para ajustar curvas com fins de calibração, entre elas, o ajuste a uma linha recta ou a polinómios de grau diferente por regressão linear por mínimos quadrados. Em qualquer caso, o programa sobrepõe à curva de ajuste os níveis de confiança a 95% e faz predição inversa (obtenção de x a partir de y) com os seus limites de confiança a 95%. Estes intervalos de confiança, permitem ao investigador não só conhecer a concentração de uma sustância em uma amostra mas também estimar a precisão da análise. No gráfico à esquerda apresenta-se a calibração de uma substância, onde os dados se ajustam a uma recta por regressão linear com pesos estatísticos. Como é sabido, a *regressão linear* usa-se quando as equações são lineares nos parâmetros (aparecem como somas ou subtracções) e se baseia em minimizar o somatório dos residuais ao quadrado. Normalmente, cada ponto é ponderado com um peso estatístico definido como o inverso do desvio padrão ($w_i = 1/s_i$), donde o desvio padrão (s_i) se pode obter a partir de réplicas. Portanto, o ajuste baseia-se em minimizar a expressão:

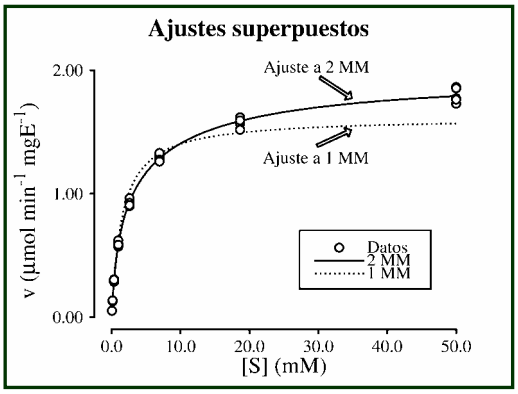
$$WSSQ = \sum (1/s_i^2)\left(y_i - f(x_i)\right)^2$$

No caso da regressão linear a solução é única e o método é exacto. Por outro lado, para avaliar a *bondade do ajuste*, utilizam-se diferentes critérios. Uns referem-se aos *residuais*: como são o valor do somatório de residuais ao quadrado, o coeficiente de correlação ao quadrado, a distribuição gráfica dos residuais (ao acaso, com quebras). Outros referem-se aos *parâmetros*: como são as variâncias dos parâmetros, seus limites de confiança a 95 %...etc.

Ajuste a funções exponenciais em sistemas cinéticos: O uso de monoexponenciais, biexponenciais, etc., é muito frequente na interpretação de muitos fenómenos cinéticos (reacções de ordem 1, reacções consecutivas, farmaco-cinética compartimental). *SIMFIT* contempla a possibilidade de ajustar diferentes tipos de exponenciais, tanto decrescentes como crescentes, e sempre com a opção de ajustar uma soma de exponenciais até ao grau desejado. Estes ajustes realizam-se sempre por *regressão não linear* para a função directa (os exponenciais são não lineares nos parâmetros), e nunca por *regressão linear* a alguma transformação linear (por exemplo logarítmica); seria necessário transformar igualmente os desvios padrão dos dados para fazer os pesos estatísticos correctamente, e isso é difícil e não é preciso. Na regressão não linear, o problema de encontrar os parâmetros óptimos é mais complexo, já que não existe uma solução única de tipo explícito (como ocorre na regressão linear), pelo que há que utilizar métodos interactivos (de procura, de gradiente), que tratam de encontrar o mínimo de WSSQ de forma aproximada.

Cinética enzimática: Imaginemos uma preparação enzimática de duas isoenzimas, com as quais se faz um estudo cinético para determinar se as 2 isoenzimas têm a mesma V_{max} y K_m ou possuem

valores diferentes, e em cada caso, estimar estes valores. Em essência, trata-se de discriminar se a equação de velocidade requer 1 ou 2 condições de Michaelis-Menten (MM). Para isto, seria necessário fazer um ajuste de regressão não linear com pesos estatísticos aos dois modelos alternativos:



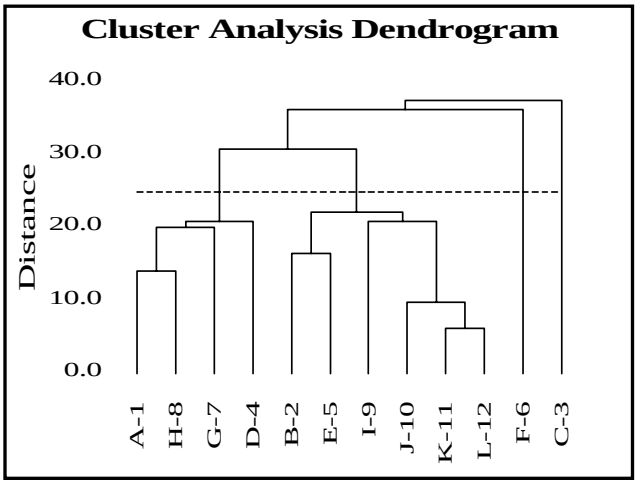
Neste exemplo, o ajuste a 2 MM é melhor que o

$$v = \frac{V_{\max_1}[S]}{K_{m_1} + [S]} \quad ; \quad v = \frac{V_{\max_1}[S]}{K_{m_1} + [S]} + \frac{V_{\max_2}[S]}{K_{m_2} + [S]}$$

ajuste a 1 MM, como se pode ver na sobreposição dos dois ajustes aos pontos experimentais (figura à esquerda). Esta confirmação visual é muito valiosa para um investigador; no entanto fazem falta outros critérios estatísticos que os determinem quantitativamente, e entre eles encontra-se o teste F, o critério de Akaike e outros.

Etapa 4. Técnicas estatísticas especiais

Para analisar certo tipo de dados existem procedimentos estatísticos mais avançados. É o caso da *análise multivariância*, no qual se dispõe de numerosas variáveis para diferentes sujeitos e se deseja extrair a máxima informação possível. Entre estas técnicas encontra-se a *"análise de conglomerados"* (análise *cluster*), que se baseia em estabelecer primeiro um tipo de medida para a distância entre casos e depois proceder ao seu agrupamento, o qual se pode representar em forma de "dendrograma" (figura da esquerda). Uma técnica semelhante é o *agrupamento a volta de "k médias"* escolhidas pelo investigador (*k-means clustering*). Outro método de análise multivariância é o chamado "*análises por componentes principais"*, que trata de transformar as variáveis originais em outras novas variáveis que expliquem a maior variabilidade possível. Deste modo, umas poucas variáveis novas permitem prescindir das muitas variáveis originais e reduzem a dimensão da análise posterior. Estas técnicas multivariantes projectam-se em muitos dos algoritmos usados em Bioinformática; por exemplo a análise de conglomerados utiliza-se em estudos dos *chips* de DNA, com o fim de identificar os genes que têm modelos similares de expressão.



Em resumo, existem muitos procedimentos de análise de dados, muitos programas e pacotes estatísticos. No entanto, é o investigador quem deve compreender os seus dados, saber os fundamentos teóricos das distintas técnicas e reconhecer qual é a mais apropriada.