

# Algumas considerações sobre dados, ajustes e incertezas

Carla Carmelo Rosa / LF3

October 10, 2021

## 1 Análise e tratamento de dados - algumas notas

A inferência científica é baseada na extrapolação de medidas a partir de dados recolhidos experimentalmente em ambiente controlado. No processo de concepção da experiência, sua concretização, e final apresentação de resultados, uma boa parte do tempo de trabalho é dedicado ao tratamento da informação gerada, e à preparação de apresentação de resultados nela baseados. Para o sucesso experimental, é importante

- boa recolha de dados;
- (pré-) processamento adequado
  - análise exploratória;
  - caracterização estatística dos dados;
  - conversão da medida em quantidades “significantes”
  - comparação/validação dos dados com modelos teóricos;
  - análise crítica de significâncias
- geração de informação gráfica de suporte (gráficos, infografias,...)

Em todos estes passos o cientista tem de ter uma postura correta, ética, de forma a produzir a informação científica mais fidedigna possível. Não pode deixar-se levar pelos seus preconceitos, o que pode levá-lo a manipular os dados, ou a sua representação, para que se moldem a uma certa “expectativa” ou presunção. Pelo contrário, deve assumir o papel de “detetive”, que procura as razões científicas ou técnicas para os desvios inesperados. Ao fim ao cabo, o erro do experimentalista, ainda que existente, terá sempre um processo físico na sua génese, que tem de ser identificado.

Nos laboratórios de ensino temos um controlo manual das experiências, e a própria aquisição de medidas é feita por observação humana na maioria das experiências. Algumas experiências indicam a conveniência de uma aquisição automática, ou pela necessidade de realizar medidas simultâneas de vários parâmetros (por exemplo, no trabalho do comportamento de um metal e um díodo, num banho térmico), ou pela rapidez do processo em estudo (como no registo da emissão fluorescente de um cristal).

Por outro lado, os tempos modernos caracterizam-se por geração constante de infinidades de dados, disponíveis para processamento! Os maiores desafios contemporâneos não são a geração

de dados, mas sim o seu conveniente tratamento e processamento. Para tal, são necessárias ferramentas adequadas, robustas, que permitam versatilidade nos métodos e na manipulação de informação.

## 1.1 WYSIWYG vs WYSIWYM, versão *análise de dados*

É importante ter noção de que não existe *A Ferramenta* de análise de dados. Existem vários pacotes de software dedicados à análise de dados, e devemos escolher aquela que melhor satisfaz o propósito do trabalho a realizar. Todas terão vantagens e desvantagens. Nos primeiros anos do curso trabalharam certamente com várias ferramentas, muito provavelmente dando grande predominância a folhas de cálculo (Excel, ou similares). O Excel é óptimo para automatizar cálculos em conjuntos limitados de dados, mas começará a apresentar limitações se quiserem fazer análises um pouco mais específicas, que possam fugir ao contexto normal de aplicação de folhas de cálculo. Por outro lado, a manipulação de conjuntos de dados de maior dimensão também se torna problemática, pois aplicações como o Excel dependem da visualização dos dados em tabela. Outras ferramentas carregam os dados em memória, associando-os a variáveis abstractas (Python, Matlab, R, SPSS, Veusz, ...). Estamos também aqui a aplicar potencialmente os paradigmas discutidos em “Comunicação em Ciência” do **WhatYouSeeIsWhatYouGet** vs **WhatYouSeeIsWhatYouMean**

## 1.2 Algumas ferramentas de análise

Existem programas gráficos (com interacção comparável àquela do EXCEL) desenvolvidos apenas com a análise de dados em mente; disponibilizam funcionalidade extremamente úteis no processamento genérico de dados (filtragens, estatística, ajustes de curvas não lineares, análise espectral por FFT ou wavelets, ...) e têm a capacidade de gerar gráficos científicos de elevada qualidade; estão também preparados para manipular conjuntos de dados de elevada dimensão, sem sobrecarregar o peso de memória gráfica envolvido na visualização destes dados na aplicação. Exemplos destes tipos de software comercial são o OriginLab, kaleidaGraph e as suas versões *OpenSoftware* SciDAVis, LabPlot, desenvolvidas tendo por inspiração os gigantes comerciais, assim como encontramos o Octave, ou o SciLab como alternativas *OpenSoftware* desenvolvidas pela comunidade científica como alternativa ao dispendioso (mas estável) Matlab.

Na UC de LF3 sugere-se a exploração do JupyterLab (Python) e/ou do SciDAVis como alternativas ao Excel para processamento de dados. Destacam-se em seguida algumas características e informações importantes para o bom funcionamento destes programas. Esta sugestão não é limitada ao âmbito de LF3, ela é feita com a convicção de que poderão ser ferramentas com elevado potencial para utilização futura académica ou profissional.

### 1.2.1 SciDAVis

O SciDAVis dá a possibilidade de gerir “projecto”, agrupando em subpastas do ficheiro .scprj diferentes entidades (tabelas, gráficos, ...) relacionadas. Pode ser útil para agrupar num único ficheiro todos os trabalhos de LF3, ou em alternativa separar dados de diferentes partes de uma experiência, e/ou condições... Usar a gosto.

[a confirmar!] o SciDAVis corre rotinas numéricas implementadas em Python, e esta linguagem utiliza o ponto (“1922.30”) como separador decimal. Pode haver impacto no bom funcionamento do programa quando o separador decimal configurado seja, por defeito, a vírgula (1922,30) utilizada na convenção portuguesa. - Forçar a que seja utilizado o ponto pode evitar erros (?)

**aviso!**- como em todo o software, é boa política gravar o ficheiro de trabalho assim que o software arranca, para permitir a gravação automática dos dados. Em ajustes SciDAVis, se por ventura algum parâmetro fica mal definido e a rotina de ajuste cai em loop fechado, o programa só terminará à força, e eventuais alterações após a última gravação serão perdidas!

### 1.2.2 Python: jupyter lab

- interface prática
  - navegação de pastas
  - visualizador de ficheiros conhecidos (.csv, imagens, ...)

## 2 Da medida ao resultado

### 2.1 Calibração dos instrumentos

A realização de medidas de uma grandeza experimental  $y$  utiliza equipamentos que desejamos em óptimo estado, e devidamente calibrados. Entende-se por instrumento calibrado o aparelho que:

- apresenta a medida correcta quando exposto a uma amostra de valor  $y_{calib}$  conhecida, i.e  $y = y_{calib}$
- apresenta resposta nula  $y_0 = 0$  quando exposto a estímulo-nulo ou na ausência de amostra.

Se não conseguirmos garantir estas condições as medidas podem ser afetadas de erros sistemáticos grosseiros que distorcem a estimativa das grandezas pretendidas. O primeiro caso pode ser contabilizado se for conhecida uma função analítica tal que

$$y_{estimado} = f(y_{medido}).$$

O segundo caso é mais comum e de fácil resolução, pois pode configurar-se como um erro sistemático  $\delta y_{sist}$  que afecta toda a experiência de igual forma (é necessário verificar se podemos mesmo assumir estes pressupostos!):

$$y_{medida} = y_{real} + \delta y_{sist}$$

Assim, todo o ensaio/experiência deve estar bem calibrado/a, e cabe ao experimentalista identificar as fontes de erros ou condicionantes sistemáticos que podem afectar a medição, anotando-as, para posterior consideração na análise de dados.

### 2.2 Parâmetros dos Ajustes Numéricos

Face ao exposto, qual a distância de trabalho entre uma lei física expressa matematicamente, e o ajuste de uma curva aos dados experimentais adquiridos no laboratório?

Escrito de outra forma, será prudente aplicar ajustes aos dados experimentais sem ponderar cuidadosamente as correspondências entre as grandezas e/ou variáveis do modelo (ajuste) e as grandezas e/ou variáveis correspondentes na medida experimental? Sem ponderar possíveis interferências no processo de medida?

Veja-se o caso de um processo de comportamento oscilatório, esperado como

$$y_{teor}(t) = A \cos(-a \times t).$$

Que pressupostos estão implícitos nesta equação que podem não ser verificados no laboratório, durante a execução da experiência?

Os dados experimentais, entre outros, podem comumente ser afetados por:

- ruído aditivo na grandeza medida ( $\delta n(t)$ );
- erros de calibração do zero do aparelho de medida ( $\delta y_{sist}$ );
- a medida no instante de tempo que consideramos inicial pode ser tal que  $|y_{med}(t=0)| \leq A$ , o que indicia que haverá um  $t_0$  tal que  $|y_{med}(t_0)| = y_{teor}(0)$ , implicando que  $t = t_{med} - t_0$ !

Resumindo, podemos traduzir estas observações e “interferências” na forma

$$y_{med}(t) = A \cos(-a \times t + \phi_0) + \delta y_{sist} + \delta n(t)$$

ou

$$y_{med}(t) = A \cos(-a \times (t - t_0)) + \delta y_{sist} + \delta n(t),$$

ligadas pela equivalência

$$\phi_0 = a \times t_0$$

que garante que o valor esperado da medida no caso particular de observação do máximo, é obtido para

$$\begin{aligned} \langle y_{med}(t_0) \rangle &= A + \delta y_{sist} \\ \langle y_{med}(0) \rangle &= A \cos(\phi_0) + \delta y_{sist} \end{aligned}$$

Face ao exposto, vimos que pode ser necessário alterar a equação modelo de forma a conter tanto o modelo físico, como as interferências associadas e identificadas na experiência, pela:

- inclusão de uma constante de ajuste, a determinar, que associamos a erros sistemáticos aditivos,  $\delta y_{sist}$ ;
- inclusão de uma constante de correção do referencial de tempo, a determinar, translacionando a origem para  $t_0$ ;
- inclusão de outros factores que possam, por razões bem identificadas, estar associados aos dados registados.

Não cuidar estes aspectos poderá traduzir-se em...:

- Os mesmos dados, utilizados de formas distintas, poderão originar ajustes (fits) semelhantes, mas de erros significativamente distintos. Ou até fits completamente diferentes!
- a definição da função de ajuste pode condicionar fortemente a qualidade do resultado obtido
- em casos mais complexos, é importante introduzir uma estimativa da ordem de grandeza dos parâmetros a determinar.

### 2.3 Ajustes: normalização de constantes

Outro dos problemas comuns em ajustes é a dificuldade de cálculo quando se misturam operações entre números de ordens de grandeza bastante diferentes.

Por exemplo, na verificação de uma lei experimental do tipo:

$$f(u) = \exp(-u),$$

em que  $u = \alpha \times x$ . Suponhamos que medimos a variável  $x$  na gama de  $10^{-6}$ . A função exponencial só produz valores *significativos*  $u$  contidos no intervalo  $[0,5]$ , pois para  $u > 5$ ,  $f(u) \rightarrow 0$ .

Qual a consequência desta observação?

... a consequência é que para podermos estimar que para valores mensuráveis de  $u$ , contidos no intervalo  $[0,5]$  e  $x \sim 10^{-6}$ , estimamos que  $\alpha \sim 10^6$ !!

esta discrepância de ordens de grandeza pode ser conseguida por **renormalização** das variáveis e constantes da equação, de forma a que os valores das constantes a ajustar sejam da ordem da unidade. Para tal, podemos usar as seguintes estratégias:

- reescrevem-se as variáveis como sendo adimensionais, introduzindo os correspondentes factores de unidade na equação; no exemplo referido, tal é obtido fazendo a transformação para a variável adimensional  $x'$ , tal que:

$$x = x' \times 10^{-6}\text{m}$$

e substituindo na equação modelo  $f(x) = f(x'10^{-6}) = f(x')$ , resultando em:

$$f(x') = \exp(-\alpha(x'10^{-6})) = \exp(-\alpha'x').$$

em que se considera  $\alpha' = \alpha 10^{-6}\text{m}$ , também adimensional.

- outra possibilidade é simplesmente definir que a análise de dados considera os dados adquiridos experimentalmente na subunidade  $[x] = \mu\text{m}$ , e que, consequentemente, o parâmetro  $\alpha$  que poderá ser determinado por ajuste numérico será expresso em unidades de  $[\alpha] = \mu\text{m}^{-1}$ .

Passámos assim para uma representação em que variáveis e constantes normalizadas são da mesma ordem de grandeza. O problema pode ser tratado numericamente neste sistema, e se necessário numa fase posterior reconverter a unidades que sejam informativas e relevante para a apresentação de informação.

### 2.4 Qualidade dos ajustes

Todos estamos habituados a procurar nos factores de covariância e de correlação associados ao ajuste pela técnica dos mínimos quadrados algum grau de certeza e confiança nos parâmetros obtidos. Sabemos que um factor  $R$  próximo de 1 é indicativo de um bom sinal, e de facto, se os dados estiverem muito próximos da recta de ajuste será fácil obter factores  $R$  perto do valor 1.

No entanto, o Método dos Mínimos Quadrados parte da assunção que a distribuição dos pontos ajustados à recta de ajuste é aleatória, e de média nula, indicando que as flutuações observadas são, à partida, de natureza estocástica. Se, por observação dos resíduos (as diferenças entre medida e ajuste) apresentar uma distribuição de pontos não aleatória (veja-se o exemplo da figura 1),

tal é fortemente indicativo de que a função de ajuste utilizada não é apropriada, e deverá ser reponderada. Tal acontece facilmente, quando a dispersão de pontos é considerável, com um desvio padrão estatístico considerável face ao valor da grandeza (por exemplo, experiência da radioatividade ..., verificação da lei de Malus, decaimento fluorescente da emissão do cristal, ...).

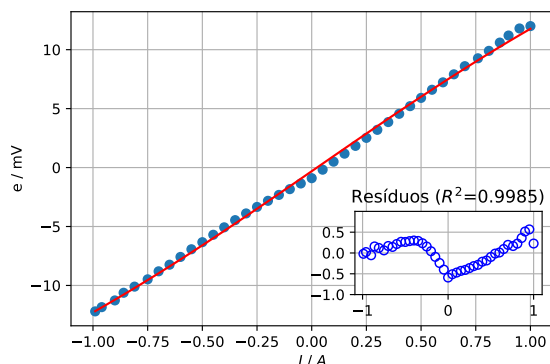


Figure 1: Ajuste linear inadequado - atente-se que o gráfico de resíduos não tem comportamento aleatório!

O gráfico de resíduos deve ser tido como uma verificação de segundo nível sobre a adequabilidade de ajustes numéricos, mesmo que bem conhecida a função de ajuste mais adequada, pois a falta de inclusão de parâmetros que possam contribuir para o modelo da medida irão manifestar-se fortemente no gráfico de resíduos.

## 2.5 Determinação experimental de resposta nula

Existem experiências onde é necessário determinar condições particulares a que corresponde uma resposta nula. Vejam-se, por exemplo, as experiências da determinação da constante de Planck, ou do estudo de materiais supercondutores. A primeira é realizada este ano (20/21) no Lab. A segunda, envolve a medida da resistência elétrica da amostra à medida que se faz variar a sua temperatura, arrefecendo-a até temperaturas criogénicas suficientemente baixas para passar ao estado supercondutor.

As estratégias para definir o “estado 0” podem depender do tipo de resposta manifestada pelo sistema. Como tal, podem envolver:

- cálculo de assíntotas nos regimes limite (incerteza experimental potencialmente maior)
- derivadas (incerteza experimental menor, critério de aferição facilmente reproduzível)
- transições pelo estado “zero” (incerteza fortemente dependente da existência de erros sistemáticos que podem mascarar ou translacionar o patamar “zero”)

### 3 Sobre incertezas, e a estimativa da incerteza final

#### 3.1 Que informação retiramos da propagação de erros nos cálculos?

Após realizar o ajuste dos pontos experimentais a uma função matemática, é importante identificar o erro final resultante das grandezas determinadas e das medidas experimentais por propagação de erros.

Quando se verifica que, dentro da precisão de medida pretendida, o erro propagado é significativo, deve ser feita uma análise termo a termo da equação de propagação de erros, para procurar identificar qual o parâmetro com maior peso na incerteza da experiência. Esta informação pode conter pistas fundamentais para a correcção ou proposta de melhoramento da experiência realizada, e é um indicador do ponto mais vulnerável da medida.

#### 3.2 Comparação do valor obtido com valor esperado

As experiências laboratoriais podem ser feitas para gerar dados novos, nunca antes medidos, ou para verificar processos conhecidos, pela medida de grandezas bem definidas.

De facto, mesmo quando são montadas novas experiências, é necessário validar os métodos e configuração experimental implementada, e tal é feito pela realização de medidas bem conhecidas, antes de explorar situações novas. Este procedimento valida o bom funcionamento e o método experimental proposto.

Para que a experiência seja validada, as medidas de quantidades conhecidas serão aferidas por comparação com os *ditos* valores esperados, ou valores teóricos, através do bem conhecido erro de exactidão.

A incerteza experimental da medida dependerá fortemente dos instrumentos, de erros sistemáticos, dos ajustes numéricos, do processamento da informação, pelo que a incerteza final da medida pode ser convertida numa precisão (incerteza experimental, face ao valor medido...) que raramente estará limitada à chamada “precisão do aparelho”. Há muitos factores a influenciar os resultados finais.

Por outro lado, a incerteza de grandezas medidas (e propagadas) que estejam caracterizadas estatisticamente é dependente do desvio padrão, e deve ser apresentada como

$$y_{\text{estimado}} = \langle y \rangle \pm 2\sigma_y,$$

o que define o intervalo dos 96% valores mais prováveis para  $y$ .

Juntando tudo, quando apresentamos um valor final para uma grandeza, será apresentado como

$$y_{\text{final}} = \langle y \rangle \pm \Delta y$$

indicando que o valor real poderá estar dentro do intervalo

$$y_{\text{real}} \in [\langle y \rangle - \Delta y, \langle y \rangle + \Delta y].$$

Na validação, é importantíssimo avaliar se este intervalo contém o valor teórico (no caso de constantes universais), ou o valor esperado (quando não temos indicação explícita do valor mas conhecemos valores indicativos, como por exemplo a resistividade de um material, dependente de

impurezas entre outras características... ). Para tal utiliza-se o erro de exatidão, ou a exatidão. Um gráfico simples com indicação do valor esperado e o(s) valor(es) obtidos experimentalmente com a(s) respetiva(s) barra(s) de erro são muito ilustrativos.

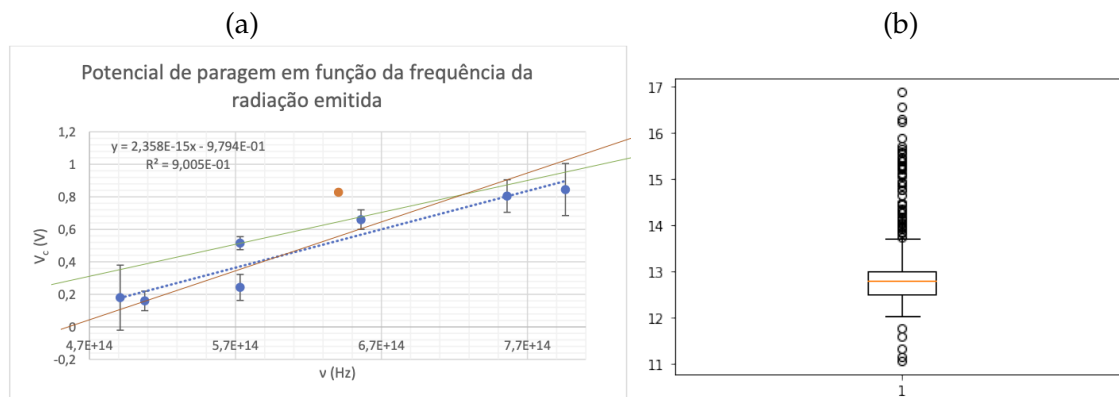


Figure 2: Exemplos de representação de medidas experimentais com respetivas incertezas, em experiências: (a) com variação de um parâmetro (frequência) e correspondentes medidas de sinal  $V$  com barra de erro - ajuste linear (azul) e estimativa de rectas de maior e menor declive (verde e vermelho); (b) com repetição da medida, resultando num valor esperado (média), correspondente incerteza e distribuição de medidas - gráfico boxplot.

Na comparação, múltiplos cenários podem surgir, e deve-se sempre discutir as implicações do resultado obtido.

Uma boa experiência será, em princípio, aquela em que o erro de exatidão é inferior à incerteza experimental. Uma experiência em que a incerteza experimental é menor que o erro de exatidão revela que potencialmente as medidas se encontram enviesadas, ou que as incertezas experimentais não foram devidamente contabilizadas.

Uma das causas mais comuns do último cenário é a incerteza experimental basear-se apenas e erradamente na propagação das incertezas associadas aos parâmetros de ajustes, esquecendo outras fontes de incerteza experimental normalmente mais condicionadoras, ou a limitações da própria medida. É importante desenvolver espírito crítico e conseguir identificar estas contribuições, pois serão estas os pontos a trabalhar para melhorar o método em uso.

[ ]: