ANÁLISE ESTATÍSTICA DE MEDIDAS EM CIÊNCIAS EXATAS

Trabalhos de meio de curso

Vito R. Vanin, Philippe Gouffon, Otaviano Helene

Março de 2017

Conteúdo

Introdução

As propriedades estatísticas dos estimadores podem ser deduzidas teoricamente em certos casos particulares. Por exemplo, a média e a estimativa da variância de dados normais têm f.d.p.s analíticas, que foram tabeladas muito antes do advento dos computadores. Essas distribuições, como várias outras obtidas analiticamente, são muito importantes porque fornecem estimativas rápidas e permitem obter resultados bastante aproximados em muitas situações, o que com frequência é tudo o que se necessita. Quando a dedução da f.d.p. da grandeza é impraticável e for importante determinar as propriedades da variável aleatória com precisão, aplicar técnicas de simulação constitui uma alternativa. Os temas destes trabalhos destinam-se a praticar essa abordagem, por meio da solução de problemas de dois tipos:

i. Comparar o resultado da simulação com o teórico em situações em que é possível determinar a f.d.p. do estimador.
ii. Determinar o comportamento de estimadores que não têm f.d.p. analítica e verificar sua adequação, o que exige a escolha de critérios para avaliá-los.

Os temas escolhidos são centrais em muitos procedimentos típicos de análise de dados, embora dificilmente um experimentador terá que enfrentá-los todos. A abordagem por seminários busca envolver mais os estudantes com a prática do tratamento de dados e provocar maior interação entre os estudantes da turma. Assim como no estudo da Física, a colega ao lado é um reservatório de conhecimento importante a quem se pode recorrer ao enfrentar um problema novo.

Os exercícios foram formatados da seguinte maneira. O primeiro parágrafo é um enunciado breve (lacônico?) do problema, em linguagem formal, que termina com a pergunta. Como muitos desses temas ainda não foram tratados na disciplina, eles podem ser difíceis de compreender e, por isso, em seguida apresenta-se a solução em etapas sucessivas, com mais detalhes. Todos os exercícios envolvem simulação, portanto serão necessários geradores de números aleatórios para resolvê-los. Embora as linguagens de computação atuais disponham de algoritmos para todas as distribuições que precisaremos na disciplina, ainda é útil entender como esses algoritmos funcionam, embora não convenha deter-se muito nesse assunto – os geradores de aleatórios dos principais programas são muito bons e cômodos de usar. O roteiro de solução da maioria dos problemas sugere testes simples dos geradores de aleatórios, a fim de enfatizar a necessidade de verificar todo programa que se usa na vida profissional – somente use um programa depois de testá-lo. O teste de um programa é também uma verificação de que você entendeu e está usando corretamente a interface por meio da qual o acessa.

Os temas abaixo que não forem objeto de um seminário serão usados como exercícios de aplicação do Mathematica, ao longo do semestre.

1. A normal como aproximação

Parte A. Funções suaves

Se *r* é uma variável aleatória com f.d.p. normal, quais as funções densidade de probabilidade do perímetro $2\pi r$, da área $4\pi r^2$ e do volume $4\pi r^3/3$? Calcule analiticamente as funções densidade de probabilidade, defina as condições limite para validade da aproximação gaussiana e teste-as (as fdps e as condições limite) por simulação. Interprete as barras de incerteza, para mais e para menos, desiguais, nos casos das grandezas área e volume.

Etapa 1. Gere números aleatórios gaussianos pelo *método da inversão* a partir de números aleatórios com f.d.p. uniforme.[1] A fim de gerar aleatórios gaussianos de média 0 e desvio padrão 1, por meio desse algoritmo, proceda da seguinte maneira:

- Construa uma tabela com os valores da integral

$$I(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

- Sorteie um número aleatório *z* entre 0 e 1 probabilidade constante nessa faixa.

- Procure *x* na tabela de $I(x)$ tal que

$$I(x) \leq z \leq I(x')$$

  com *x* e x' entradas consecutivas na tabela.

- O aleatório gaussiano μ pode ser obtido por interpolação linear,

$$\mu = x + \frac{x' - x}{I(x') - I(x)}\big(z - I(x)\big)$$

  Não é necessário montar uma tabela de *I*(*x*) com mais que 100 valores.

- Finalize esta etapa verificando se os aleatórios são gaussianos mesmo. Para tanto, sorteie um número grande de aleatórios gaussianos, calcule a média, o desvio padrão, a curtose, κ,

$$\kappa = \frac{\overline{(x - \bar{x})^4}}{(\sigma^2)^2} - 3$$

  e o parâmetro de assimetria, β,

[1] Embora as linguagens de computação atuais gerem números aleatórios com distribuição normal, cumpra todos os passos desta etapa, cujo objetivo é apresentar o método da inversão, em que se baseiam muitos geradores de aleatórios. Além disso, certas funções de probabilidade permitem inversão analítica, o que pode aumentar a velocidade de cálculo em simulações muito longas.

$$\beta \quad = \frac{\overline{(x-\bar{x})^3}}{(\sigma^2)^{3/2}}$$

Etapa 2. Supondo a variável r gaussiana de média r_0 e desvio padrão σ_r (conhecidos), calcule as fdps da área $A = 4\pi r^2$ e do volume $V = 4\pi r^3/3$.

- Calcule os valores médios de A e de V, A_0 e V_0, em função de r_0 e σ_r.
- Expanda as fdps de A e V em torno de A_0 e V_0 nos casos em que

$$A - A_0 \ll A_0,$$

$$V - V_0 \ll V_0,$$

  obtendo aproximações gaussianas para essas fdps.
- Compare as funções exatas com as aproximadas nos pontos: $A = A_0$; $A = A_0 \pm \sigma_A$; $A = A_0 \pm 2\sigma_A$ e nos pontos equivalentes para V (use σ_A calculado aproximadamente).
- Qual o valor de σ_r/r que origina diferenças de 5% entre os valores das fdps exata e a aproximada nesses pontos escolhidos?

Etapa 3. Escolha r_0 e σ_r para os quais as aproximações gaussianas das fdps de A e V sejam válidas, gere aleatórios gaussianos com essa média e desvio padrão – esse é o conjunto $\mathbf{R}_1 = \{r_i\}$, (valores simulados com distribuição noramal de média r_0 e desvio padrão σ_r). Construa os conjuntos $\mathbf{A}_1 = \{A(r_i)\}$ e $\mathbf{V}_1 = \{V(r_i)\}$. Calcule a média, o desvio padrão, a curtose e a assimetria dos 3 conjuntos.

- Repita toda a operação acima para r_0 e σ_0 tais que as aproximações gaussianas das fdps de A e V *não* sejam válidas. Neste caso, escolha os intervalos de A e V que melhor representam as grandezas – discuta as possibilidades e defenda sua escolha!

Parte B. Propagação da incerteza em um expoente

Se t é uma variável aleatória gaussiana, como se distribui $f(t) = \exp(\lambda t)$? Calcule analiticamente a função densidade de probabilidade, defina as condições limite para validade da aproximação gaussiana e teste-as (a f.d.p. e as condições limite) por simulação. Interprete as barras de incerteza assimétricas, σ_{f+} e σ_{f-}.

Etapa 4. Supondo a variável t gaussiana de média t_0 e desvio padrão σ_t (conhecidos), calcule a f.d.p. de $f(t) = \exp(\lambda t)$.

- Calcule o valor médio de f, f_0, em função de $f(t_0)$ e σ_f.
- Expanda a f.d.p. de f em torno de f_0 no caso em que

$$f - f_0 \ll f_0,$$

  e obtenha uma aproximação gaussiana para essa f.d.p..
- Compare a função exata com a aproximada nos pontos: $f = f_0$; $f = f_0 + \sigma_f$; $f = f_0 - \sigma_f$ (use σ_f calculado aproximadamente).

- Qual o valor de σ_t que origina diferenças de 5% entre os valores da f.d.p. exata e a aproximada nesses pontos escolhidos?

Etapa 5. Escolha t_0 e σ_t para os quais a aproximação gaussiana da f.d.p. de f seja válida, gere aleatórios gaussianos com essa média e desvio padrão – $\mathbf{T}_1 = \{t_i\}$ (aleatórios gaussianos de média t_0 e desvio padrão σ_t). Construa o conjunto $\mathbf{F}_1 = \{f(t_i)\}$. Calcule a média, o desvio padrão, a curtose e a assimetria de cada um dos 2 conjuntos.

- Repita toda a operação acima para t_0 e σ_t tais que a aproximação gaussiana da f.d.p. de f não seja válida. Nesse caso, escolha o intervalo de f que melhor representa as grandezas — discuta as possibilidades e defenda sua escolha!

Etapa 6. Compare a condição de validade da aproximação gaussiana nesse caso com aquela de transformações do tipo t^n (Parte A) e discuta qual o motivo da diferença de critérios.

2. As incertezas associadas às interpolações e extrapolações

Calcule o valor interpolado ou extrapolado a partir de um conjunto de pontos { (x_i,y_i,σ_i) } onde *y* é uma função linear nos parâmetros e em *x* também. Não esqueça de incluir a covariância entre os parâmetros da reta ajustada na propagação das variâncias! Verifique, por simulação, que as estimativas das variâncias são adequadas. Você pode tentar descobrir o que acontece se agregar um dado interpolado ao conjunto de dados originais.

Etapa 1. Gere aleatórios gaussianos com o algoritmo de Box-Muller**:**

- Gere **2** números aleatórios U_1 e U_2, com f.d.p. uniforme na faixa [0,1], com o gerador do programa que estiver usando.
- Calcule
  - $X_1 = \sqrt{-2 \ln U_1} \cos 2\pi U_2$
  - $X_2 = \sqrt{-2 \ln U_1} \operatorname{sen} 2\pi U_2$

  onde X_1 e X_2 são dois números aleatórios gaussianos independentes de média 0 e desvio padrão 1.
- Repita as operações acima até obter muitos aleatórios gaussianos.
- Verifique se os números obtidos são mesmo aleatórios gaussianos: calcule a média e o desvio padrão; faça um histograma e verifique se as quantidades de números nas faixas $[-3\sigma,-2\sigma]$, $[-2\sigma,-\sigma]$, $[-\sigma,0]$, etc, obedecem à fdp normal; calcule os parâmetros de assimetria e curtose e veja se coincidem com os valores esperados, dentro das incertezas.
- Calcule a estimativa da cov(X_1,X_2) e verifique se de fato é nula dentro da incerteza dessa estimativa.
- A partir da hipótese que U_1 e U_2 têm f.d.p. uniforme na faixa [0,1], mostre analiticamente que X_1 e X_2 tem fdps normais. Transforme as variáveis aleatórias como discutido no capítulo 2 do livro.

Etapa 2. Gere um conjunto de pontos simulados { (x_i,y_i,σ_i) }, com *x* a variável independente e *y* a variável dependente, e

$$y_i \approx N(a_0 + b_0 x_i, \sigma_i) \quad ,$$

em que ≈ significa "distribuído conforme" a f.d.p. que segue o símbolo e em que a_0 e b_0 são constantes conhecidas, com σ_i igual ao desvio padrão. Vamos, então, supor que o erro da medida, ϵ_i , seja distribuído normalmente com

$$\sigma_i^2 = < \epsilon_i^2 > \quad ,$$

com a propriedade de que não há erro sistemático,

$$< \epsilon_i >= 0 \quad .$$

Use $\{x_i\} = \{4,8,12,16,20,24\}$ e $\sigma_i = 0{,}30$ para todo i, com $a_0 = -2{,}5$ e $b_0 = 0{,}225$. Escreva o algoritmo de ajuste (tão genericamente quanto possível), ajuste os parâmetros ao conjunto de dados e calcule χ^2. Com $\hat{a}$ e $\hat{b}$ obtidos, calcule os valores interpolados de y para $x = 10$, 14 e 20 e os valores extrapolados para $x = 0$ e $x = 40$, bem como os desvios padrões associados. Interprete o resultado (extrapolação *boa* ou *ruim*, em função do valor obtido conter, dentro da incerteza, o valor *verdadeiro*).

Etapa 3. Repita a etapa 2 muitas vezes. Faça histogramas dos valores de y interpolados e extrapolados e interprete o resultado. Os desvios padrões calculados são adequados?

Etapa 4. Faça um histograma dos χ^2 e interprete. Verifique se a um valor de χ^2 próximo ao médio está associado um valor interpolado próximo ao verdadeiro, fazendo diagramas de pontos ($y_{interpolado}$,χ^2). Escolha duas posições x distintas para calcular os $y_{interpolado}$ e construir esses *scatter plots*. Calcule as covariâncias entre χ^2 e os parâmetros escolhidos a partir dos resultados numéricos que você graficou.

Etapa 5. Calcule a matriz de covariância entre os dados e os valores interpolados para $x = 10$, 14 e 20.

Etapa 6 (Opcional). Ajuste a reta a esse conjunto todo (agora, com três dados a mais e com a matriz de covariância no meio) e compare os resultados obtidos com e sem a inclusão dos dados interpolados, tanto para os coeficientes quanto para o χ^2.

3. Aumentar o número de parâmetros não implica em melhorar o ajuste.

Simular uma medida $\{(x_i,y_i,\sigma_i)\}$, onde x_i é a variável independente supostamente observada sem erro e y_i é a variável dependente que se relaciona com x_i por meio da expressão

$$y = h(x;\vec{a}) + \epsilon$$

onde $\vec{a}$ é um vetor de v constantes, y depende linearmente dos a_i e ϵ é um erro não sistemático distribuído de acordo com uma Normal de desvio padrão σ conhecido. Ao final do trabalho, você deve escolher uma forma para a função h que possa ser útil nas suas pesquisas, mas comece com apenas dois parâmetros e uma fórmula simples,

$$y = a_0 + b_0 x + \epsilon,$$

onde a_0 e b_0 são constantes. Ajustar pelos dados os parâmetros de reta, parábola e cúbica, calculando χ^2 e interpretando os valores obtidos para os coeficientes parabólico e cúbico e o significado de ajustá-los. Testar a interpretação repetindo a simulação muitas vezes.

Etapa 1. Teste se os números aleatórios gaussianos do programa que está usando têm as propriedades desejadas. Gere muito números, calcule a média e o desvio padrão; faça um histograma e verifique se as quantidades de números nas faixas $[-3\sigma,-2\sigma]$, $[-2\sigma,-\sigma]$, $[-\sigma,0]$, etc, obedecem à fdp normal; calcule os parâmetros de assimetria e curtose e verifique se tem valores iguais aos esperados, dentro das respectivas incertezas. Divida os números em pares (g_1,g_2), calcule a estimativa da $\mathrm{cov}(g_1,g_2)$ e verifique se ela é nula dentro da incerteza dessa estimativa.

Etapa 2. Obtenha um conjunto de dados simulados. É preciso, primeiro, fixar os valores da variável independente, os valores verdadeiros dos parâmetros e dos desvios padrões "experimentais". Façamos as seguintes escolhas:

$$\{x_i\} = \{4,6,8,10,12,14,16,18,20,22,24\} \qquad \text{(portanto 11 dados)}$$

$$a_0 = -2{,}5 \qquad \text{e } b_0 = 0{,}225 \qquad ;$$

$$\sigma_i = 0{,}30 \text{ para todo } i \ (i = 1,2,...N, \text{ com } N = 11)$$

Para simular o conjunto de dados, obtenha, para cada um dos N valores da variável independente x_i, um número aleatório gaussiano de média $a_0 + b_0 x_i$ com o desvio padrão escolhido.

Escreva o algoritmo de ajuste dos parâmetros das funções $y = a + bx$ e $y = c + dx + ex^2$ (tão genericamente quanto possível) e obtenha a_1, b_1, c_1, d_1 e e_1, os valores ajustados dos parâmetros. Calcule χ^2 para cada um dos ajustes e interprete os resultados (ajuste *bom* ou *ruim*).

Etapa 3. Repita os cálculos com muitos conjuntos de valores simulados. Faça os histogramas dos coeficientes polinomiais obtidos, bem como dos χ^2 — claro, não misture dados de ajustes de polinômios de graus diferentes. Que acontece aos valores ajustados do termo parabólico? E aos conjuntos dos χ^2 obtidos dos ajustes de parábola?

Etapa 4. Para cada conjunto de valores simulados, verifique se $P(\chi^2 > \chi^2_{calculado})$ cresce ou decresce com o grau do polinômio, onde *P* é a probabilidade de $\chi^2_{calculado}$ ser excedido. Procure estabelecer um critério para a escolha do grau. Em que fração das simulações seu critério foi adequado? Discuta com o professor ou a professora a adequação do critério escolhido, bem antes de dar o seminário, para verificar se fez uma das escolhas comuns ou não.

Etapa 5. Verifique se há alguma dependência entre o ajuste ser *bom*, no sentido de obter-se um χ^2 próximo do número de graus de liberdade, e calcular-se um valor *nulo* para o coeficiente parabólico, no sentido do valor ajustado ser menor que o desvio padrão. Utilize, por exemplo, um diagrama de todos pontos (χ^2,coeficiente) obtidos.

Etapa 6. Adapte as etapas acima para a função $h(x; a, b, c) = a\,\mathrm{Erf}(b\,x)$+c em que Erf é a função erro. Escolha valores "verdadeiros" (*a*,*b*,c) e use-os em todas as simulações.

Etapa 7. Faça um estudo com um conjunto de dados tal que os valores de *x* cubram toda a região [0; 2/b], e verifique se os parâmetros são estimados sem tendenciosidade.

Etapa 8. Faça um estudo com um conjunto de dados tal que os valores de *x* cubram apenas uma pequena parte da região [0; 2/b], por exemplo, [1/b; 1,1/b] e verifique se os parâmetros são estimados sem tendenciosidade.

4. Ajuste de parâmetros a dados com erros nas variáveis dependente e independente.

Simular um conjunto de dados $\{(x_i,\sigma_{xi},y_i,\sigma_{yi})\}$ onde x é a variável independente observada com desvio padrão σ_x e y a variável dependente observada com desvio padrão σ_y. A função verdadeira que relaciona y e x é

$$y = a_0 + b_0x + c_0x^2$$

e os erros em x e y não são sistemáticos e distribuem-se como gaussianas independentes. Ajustar os parâmetros da parábola e verificar se o resultado é adequado. Repetir a simulação muitas vezes de maneira a obter-se histogramas dos parâmetros ajustados e verificar se esses histogramas correspondem a gaussianas.

Etapa 1. Simule um conjunto de 9 dados:

• Para $x_{i,0} \in \{-20,-15,-10,-5,0,5,10,15,20\}$, sorteie um aleatório gaussiano x_i com média $x_{i,0}$ e desvio-padrão

$$\sigma_{x,i} = \begin{cases} 1,0 & , x_i \leq 0 \\ 0,6 & , x_i > 0 \end{cases}$$

e um aleatório gaussiano de média $y_i = a_0 + b_0x_{i,0} + b_0{x_{i,0}}^2$ e desvio-padrão $\sigma_{y,i}$

Escolha $a_0 = -2,5$; $b_0 = 0,55$; $c_0 = 0,05$ e $\sigma_{y,i} = 1,1$ para todo i.

Etapa 2. Ajuste os parâmetros da parábola ao conjunto de valores sorteados $\{(x_i, y_i)\}$, levando em conta a incerteza na variável independente através de uma projeção na direção y:

- Ajuste primeiro os parâmetros da parábola *ignorando* a incerteza em x. Obtenha, então, estimativas a^0, b^0 e c^0 dos parâmetros da função.
- Calcule, para cada i, o desvio padrão efetivo σ_i como

$$\sigma_i^2 = \sigma_{y,i}^2 + \left(\left.\frac{\partial y}{\partial x}\right|_{x_i}\right)^2 \sigma_{x,i}^2$$

- Ajuste os parâmetros da parábola ao conjunto *equivalente* de dados, $\{ (x_iy_i\sigma_i) \}$ e calcule χ^2.
- Repita o cálculo de σ_i^2 acima com os novos valores dos coeficientes e refaça o ajuste até que as variações nos parâmetros e no χ^2 sejam desprezíveis.

Etapa 3. Refaça a simulação muitas vezes e faça os histogramas dos parâmetros a, b, c e dos χ^2 obtidos. Para cada parâmetro, verifique se há desvio sistemático calculando a média; se o desvio padrão calculado a partir da matriz de covariância corresponde ao calculado

a partir dos histogramas; se 68% dos valores ajustados estão a menos de um desvio padrão da média, 95% a menos de dois desvios padrões, etc.

Etapa 4. Há situações em que esse procedimento é inadequado, por exemplo se c_0 é muito grande, tal como $c_0 = 2$, ou se σ_x é grande. Por que? Escolha uma situação em que o resultado é tendencioso e simule, mostrando essa tendenciosidade. Seria possível corrigir o erro *sistemático* do ajuste, nesses casos? Como?

Etapa 5. O método dos mínimos quadrados é *inconsistente* quando σ_x é maior ou da ordem da largura da região onde estão os dados. Assim, repita a simulação com σ_x cada vez maior, até verificar que os parâmetros ajustados são inconsistentes. Procure determinar uma regra empírica para a ordem de grandeza desse valor limite.

5. Propriedades estatísticas do desvio padrão.

Parte A. Correlações entre os momentos da função densidade de probabilidade.

Simular cerca de 100 medidas, cada uma com 10 dados que obedeçam à fdp normal de média x_0 e desvio padrão σ_0. Verificar se $\bar{x}$ e σ_m são correlacionados, onde σ_m é o desvio padrão da média. Verificar que a variância da média, σ_m^2 , tem a mesma fdp que a variável χ^2. Estimar o desvio padrão do desvio padrão.

Etapa 1. Simule as medidas com 10 dados usando aleatórios gaussianos de média x_0 e desvio padrão σ_0.

Etapa 2. Para cada uma das medidas simuladas, calcule as grandezas: média, desvio padrão, desvio padrão da média, parâmetro de assimetria β, curtose κ e *range*, e faça um histograma de cada uma delas. Define-se

$$\beta = \frac{\overline{(x-\bar{x})^3}}{(\sigma^2)^{3/2}}$$

$$\kappa = \frac{\overline{(x-\bar{x})^4}}{(\sigma^2)^2} - 3$$

$$range = x_{max} - x_{min}.$$

Etapa 3. Desenhe um diagrama com eixos $\bar{x}$ e σ_m, onde coloca-se um ponto para cada par dessas grandezas obtido das medidas simuladas. Calcule a correlação ρ entre $\bar{x}$ e σ_m,

$$\rho = \frac{< (\bar{x} - \overline{\bar{x}})(\sigma_m - \bar{\sigma}_m) >}{\sigma_{\bar{x}}\sigma_{\sigma_m}}$$

onde $< a >$ significa o mesmo que $\bar{a}$, a média de a. Como você explica a inexistência da correlação entre $\bar{x}$ e σ_m? O fato dessas duas estatísticas serem funções dos mesmos dados não devia implicar numa correlação entre elas?

Etapa 4. O mesmo que na etapa 3, para β, κ e *range*.

Parte B. Desvio padrão efetivo e intervalo de confiança.

Considere que as observações de uma grandeza física x seguem uma distribuição normal. Simular muitas medidas, com 3 dados cada uma, dessa grandeza. Determinar os menores intervalos que contém o valor verdadeiro em 68% e 95% das simulações. Repetir a simulação e a determinação dos intervalos para medidas com 10 dados. Entender porque histogramas *provavelmente alargam* com o aumento do número de dados, embora tenham, em *média*, a largura seja *independente* do número de dados.

Etapa 5. Fixando os valores de x_0 (média) e σ_0 (desvio padrão), obtenha os conjuntos de 3 dados simulados.

Se quiser, use $x_0 = 20{,}0$ e $\sigma_0 = 2{,}5$, mas sinta-se à vontade para fazer *sua* escolha. Para cada conjunto, calcule $\bar{x}$, σ e σ_m. Faça histogramas dessas estatísticas. Determine em que fração dos conjuntos $x_0 \in [\bar{x}-\sigma_m, \bar{x}+\sigma_m]$ e em que fração dos conjuntos $x_0 \in [\bar{x} - 2\sigma_m, \bar{x} + 2\sigma_m]$.

Etapa 3. Determine α_{68} e α_{95} tais que as frações dos conjuntos em que $x_0 \in [\bar{x} - \alpha_{68}\sigma_m, \bar{x} + \alpha_{68}\sigma_m]$ e $x_0 \in [\bar{x} - \alpha_{95}\sigma_m, \bar{x} + \alpha_{95}\sigma_m]$ sejam 68,3% e 95,5%, respectivamente. Compare os valores de α obtidos com aqueles calculados a partir da distribuição de *t de Student.*

Etapa 4. Repita a etapa 2, agora com conjuntos de 10 dados. Compare as dispersões, valores médios, valores prováveis e medianas dos σ obtidos nas séries de simulações das Etapas 2 e 3. Compare as mesmas estatísticas calculadas agora sobre as variâncias σ^2 obtidas. O que você diz a respeito da dependência da largura média dos histogramas com o número de dados? E da largura provável?

Etapa 5. Simule uma medida com 2 dados e calcule σ. Sorteie mais um dado e *recalcule* σ. Vá acrescentando dados à medida e recalculando σ. (Desculpe a parecença com uma receita de bolo, mas uma explicação mais resumida pode pôr tudo a desandar...). Faça um gráfico de σ contra o número de dados. Repita a operação diversas vezes. Qual o comportamento mais comum – o histograma alarga, estreita ou não muda de largura?

Etapa 6. Proceda como na etapa 5, mas para o valor médio da variável $t_{\nu=1}$, que é $(x_1 + x_2)/(x_1 - x_2)$, ou seja, é a variável *t* para uma medida com 2 dados de uma grandeza de média nula e desvio padrão desconhecido. Como evolui a média dessa grandeza? e a mediana?

6. Estimativa da incerteza proveniente dos parâmetros não ajustados.

Ajustam-se os parâmetros de uma função

$$y = f(\vec{x}; \vec{a}, \vec{p})$$

a um conjunto de dados $\{(x_i,y_i,\sigma_i)\}$ onde x representa a variável independente, $\vec{a}$ o conjunto de parâmetros ajustáveis e $\vec{p}$ os parâmetros conhecidos, que têm uma precisão limitada, mas não serão ajustados aos dados. Deseja-se propagar a incerteza dos p_j no resultado final (os parâmetros a_k). Vamos representar o desvio padrão no parâmetro p_j por s_j. Vamos parar aqui a descrição genérica do problema e passar a um exemplo específico. ATENCÃO, você pode sugerir *outra* função f que lhe seja de interesse!

EXEMPLO. (Pode ser usado para o trabalho, mas seu uso não é obrigatório.)

Suponha que deseja-se separar as atividades de dois nuclídeos observadas numa mesma linha espectral, através da separação das meias-vidas envolvidas. A equação que dá o número de átomos ativos no instante t é

$$N(t) = N_a \exp(-\lambda_a t) + N_b \exp(-\lambda_b t) \quad ,$$

de maneira que o número de contagens no intervalo $(t_i,t_i + \Delta t_i)$ é

$$y(\vec{x}; \vec{N}, \vec{\lambda}) = N_a \exp(-\lambda_a t_i)[1-\exp(-\lambda_a \Delta t_i)]+N_b \exp(-\lambda_b t_i)[1-\exp(-\lambda_b \Delta t_i)]$$

onde $x_i = (t_i, \Delta t_i)$, $\vec{N}$= (N_a,N_b) e $\vec{\lambda}$ = (λ_a,λ_b). (Neste caso, o número de parâmetros ajustáveis é igual ao de parâmetros fixos conhecidos, o que não é sempre necessário.) Assim, a partir da medida $\{(t_i,\Delta t_i,y_i,\sigma_i)\}$, onde σ_i é o desvio padrão de y_i, pode-se estimar N_a e N_b por uma regressão, se λ_a e λ_b são conhecidos. A matriz de covariância do ajuste, porém, reflete apenas as incertezas dos y_i mas não as incertezas de λ_a e λ_b.

Etapa 1. Para o conjunto de dados

t(h)	Δt(h)	$y(\sigma)$
0,5	1,5	23590(159)
2,2	1,5	7157(94)
3,9	1,5	3267(69)
5,5	3,0	3180(79)
8,7	4,0	1914(77)

$$\lambda_a = 0{,}203(9) h^{-1}$$

$$\lambda_b = 0{,}920(18) h^{-1}$$

calcule $N_a(\sigma_a)$ e $N_b(\sigma_b)$.

Há dois procedimentos distintos para incluir na variância do resultado final as parcelas devidas às variâncias desses dois parâmetros, descritos nas duas secções a seguir.

Parte A. Propagação do desvio padrão por estimativas das derivadas.

Etapa 2. Estime $\frac{\partial N_a}{\partial \lambda_a}$ refazendo o cálculo de N_a com

$$\lambda'_a = \lambda_a + \Delta.$$

Use Δ pequeno, $\Delta \sim 0{,}0001h^{-1}$, e estime

$$\frac{\partial N_a}{\partial \lambda_a} \cong \frac{N'_a - N_a}{\Delta}$$

Faça o mesmo para estimar $\frac{\partial N_a}{\partial \lambda_b}$, $\frac{\partial N_b}{\partial \lambda_a}$ e $\frac{\partial N_b}{\partial \lambda_b}$

Etapa 3. Calcule então a variância total de N_a, Σ^2_a, a partir da superposição da variância devida aos y_i, σ_a^2, com a devida aos parâmetros λ,

$$\Sigma_a^2 = \sigma_a^2 + \left(\frac{\partial N_a}{\partial \lambda_a}\right)^2 s_a^2 + \left(\frac{\partial N_a}{\partial \lambda_b}\right)^2 s_b^2,$$

com uma fórmula análoga para a variância total de N_b.

Parte B. Pelo Método de Monte Carlo.

Etapa 4. Repita a etapa 1 diversas vezes, com valores sorteados de λ_a e λ_b , de acordo com gaussianas de médias e variâncias iguais aos valores conhecidos dessas grandezas. Faça um histograma dos N_a e dos N_b obtidos e calcule os desvios padrões correspondentes, $\sigma_{a(\lambda)}$ e $\sigma_{b(\lambda)}$. A variância total em N_a será

$$\Sigma_a^2 = \sigma_a^2 + \sigma_{a(\lambda)}^2 \quad ,$$

com uma fórmula análoga para a variância de N_b.

7. Fórmula de cálculo da variância de uma função de variáveis correlacionadas.

A variância de uma função y de m variáveis aleatórias a_1, a_2, ..., a_m, $y = f(a_1,a_2,...a_m)$, pode ser estimada como

$$\sigma^2 = \sum_{k=1}^{m} \left(\frac{\partial y}{\partial a_k}\right)_{\hat{a}}^2 \sigma_k^2 + \sum_{k\neq j}^{m} \left(\frac{\partial y}{\partial a_k}\frac{\partial y}{\partial a_j}\right)_{\hat{a}} \mathrm{cov}(a_j, a_k)$$

onde σ_k^2 é a variância de a_k e cov(a_j,a_k) é a covariância entre a_j e a_k. Teste, através de simulação, a adequação da fórmula para as funções:

$$y = a + b; \qquad y = a - \mathrm{b};$$

$$y = a^2\, b^2; \qquad y = a\ \mathrm{atan}(b),$$

$$y = a\ \exp(0{,}7 * b).$$

Você pode usar a = 2,00(4), b = 1,000(5), com cov(a,b) = 1,6×10^{-4}.

Etapa 1. Construa um gerador de números aleatórios gaussianos correlacionados, com o seguinte algoritmo:

- Gere dois números aleatórios gaussianos X e Y de média nula e desvio padrão 1.
- Obtenha dois números aleatórios a e b com fdp gaussiana, com desvios padrões σ_a e σ_b e coeficiente de correlação ρ, calculando

$$a = a_0 + X\sigma_a$$

$$b = b_0 + \sigma_b(\rho X + (1 - \rho^2)Y)$$

Mostre, analiticamente, que as variáveis a, b têm correlação igual a ρ.

Etapa 2. Sorteie um grande número de pares (a,b), talvez 1000. Calcule as funções y dadas, faça os histogramas dos y obtidos e calcule as variâncias a partir dos dados histogramados. Confira com o esperado da fórmula de cálculo da variância de uma função de variáveis aleatórias.

Etapa 3. Invente uma função de muitas variáveis[1] e uma matriz de covariância para as grandezas que entram nessa função, de modo que os coeficientes de correlação sejam variados, com módulos entre 0,1 e 0,9. Sorteie números aleatórios gaussianos correlacionados de acordo com essa matriz inventada, e proceda como na etapa 2. Determine quantos sorteios são necessários para estimar corretamente um coeficiente de correlação igual a 0,9 por este método, e quantos são necessários se esse coeficiente for igual a 0,1.

[1] De preferência, escolha uma fórmula com que você vai trabalhar ou já trabalhou. Qualquer número de parâmetros maior que 2 será suficiente para o problema, mas com 5 ou 6, o exercício poderá ficar mais interessante.

8. Eficiência de estimadores de posição

Três estimadores para a localização de uma distribuição são usados frequentemente: a *média*, a *mediana* e a *média dos extremos*. Neste trabalho, estes três estimadores são usados para avaliar o parâmetro de localização das distribuições: gaussiana, Cauchy e uniforme.

Um estimador é tanto mais eficiente quanto menor for sua variância. Portanto, para achar a eficiência, deve-se determinar a variância do estimador para a distribuição em estudo.

Etapa 1. Para cada distribuição, gere 100 sequências de 10 dados. Para cada sequência, calcule as três estimativas e histograme-as.

Etapa 2. Calcule a variância dos histogramas da etapa 2. Monte uma tabela com os resultados. Que estimador é o mais eficiente para cada distribuição?

Etapa 3. Repita as etapas 1 e 2 com sequências de $N = 1000$ dados.

9. Análise de previsão.

Como a matriz de covariâncias depende apenas dos valores da variável independente e das incertezas experimentais, mas não dos valores da variável dependente, é possível escolher previamente para quais valores da variável independente devem ser efetuadas as observações experimentais de maneira a minimizar-se a incerteza do resultado final. Este tipo de estudo é chamado análise de previsão.

Suponha que deseja-se separar as atividades de dois nuclídeos observadas numa mesma linha espectral, através da separação das meias-vidas envolvidas. A equação que dá o número de átomos ativos no instante t é

$$N(t) = N_a \exp(-\lambda_a t) + N_b \exp(-\lambda_b t) \quad ,$$

de maneira que o número de contagens no intervalo $(t_i,\ t_i + \Delta t_i)$ é

$$A_i = N_a \exp(-\lambda_a t_i)[1 - \exp(-\lambda_a \Delta t_i)] + N_b \exp(-\lambda_b t_i)[1 - \exp(-\lambda_b \Delta t_i)].$$

Assim, a partir da medida $\{(t_i,\Delta t_i,A_i,\sigma_i)\}$, onde σ_i é o desvio padrão em A_i, pode-se estimar N_a e N_b por uma regressão linear, se λ_a e λ_b são conhecidos. Fazendo-se um modelo para σ_i, por exemplo,

$$\sigma_i^2 = A_i + B\Delta t_i \quad ,$$

onde B é uma constante conhecida (no caso, o número de contagens de fundo sob o pico por unidade de tempo), é possível construir uma matriz de covariâncias para cada escolha dos intervalos de contagem $(t_i,\ t_i + \Delta t_i)$ e verificar qual deles resulta numa menor incerteza final em N_a e N_b. Claro, é preciso ter ideia da razão entre N_a e N_b para estimar as incertezas.

Etapa 1. Façamos a escolha das constantes envolvidas:

$$\lambda_a = 0{,}203\ h^{-1}\ \lambda_b = 0{,}920\ h^{-1}$$

$$N_a = 2N_b$$

$$N_b = 10000$$

$$B = 1000$$

M = número de intervalos de contagem utilizados = 5

Etapa 2. Monte o algoritmo para estimar N_a e N_b a partir de um conjunto de dados $\{(t_i,\Delta t_i,A_i,\sigma_i)\}$. Prepare um programa para calcular a matriz de covariâncias.

Etapa 3. Para t_i (em hora) $\in \{0,2,4,6,8\}$ e Δt_i = 1,9 h, estime σ_i da maneira sugerida no enunciado do problema e use o algoritmo da etapa 2 para calcular σ_a e σ_b.

Etapa 4. Tente minimizar σ_a (ignorando σ_b) por tentativa e erro, escolhendo outros conjuntos de pares de variáveis independentes t_i, Δt_i. Faça o mesmo com respeito ao parâmetro N_b, isto é, minimize σ_b ignorando σ_a. É possível minimizar simultaneamente σ_a e σ_b ?

Etapa 5. Faça a análise de previsão para uma situação real do seu experimento.

10. Unicidade do estimador de Mínimos Quadrados.

É comum repetirmos medidas muitas vezes, a fim de melhorar a precisão do resultado ou verificar sua repetitividade. Nesse caso, cabe se perguntar se faz diferença ajustar os parâmetros pelo Método dos Mínimos Quadrados a todos os conjuntos de dados de uma vez ou se é melhor ajustar os parâmetros a cada conjunto e tirar a média dos resultados parciais. O objetivo deste trabalho é ilustrar que, por causa da unicidade do estimador de Mínimos Quadrados, os detalhes do procedimento são irrelevantes, desde que toda a informação necessária seja usada, em particular as matrizes de covariância completas – as variâncias não bastam.

Etapa 1. Considere que, na medida do decaimento de um conjunto de átomos, a intensidade I da linha espectral em função do tempo t é

$$I(t) = I_0 \exp(-\lambda_0 t) \quad .$$

A medida consiste em amostrar a intensidade I_i em um conjunto de instantes t_i, o que dá origem a um conjunto de dados $\{(t_i, I_i, \sigma_i), i = 1..N\}$, onde σ_i é o desvio-padrão de I_i; adotaremos que os valores I_i são estatisticamente independentes.

A fim de simular o processo, necessitamos valores para os parâmetros e os instantes de amostragem. Se você conhecer valores típicos no seu ramo de pesquisa, adote-os; caso contrário, use $I_0 = 1000$ ou 2000; $\lambda_0 = 0{,}4$ ms^{-1}; $t_i = 2$, 4, 6, 8 e 10 ms e $\sigma_i =$ 40 ou 60, conforme o valor de I_0, mas igual em todos os instantes.

Sorteie *dois* conjuntos de dados, um para cada um dos valores de I_0, em que cada ponto $I_i(t_i)$ é sorteado com média $I_0 \exp(-\lambda_0 t_i)$ e desvio-padrão σ_i.

37

Etapa 2. Determine I e λ pelo MMQ para cada um dos dois conjuntos de dados. Fica mais fácil linearizar a expressão, ou seja, ajuste os parâmetros a ln I. Determine a matriz de covariâncias das estimativas e o valor de χ^2 em cada um dos ajustes.

Etapa 3. Determine a soma dos χ^2 e as médias dos dois valores de λ obtidos na etapa anterior.

Etapa 4. Ajuste os parâmetros I, I_0 e λ, considerando todos os 10 pontos experimentais simultaneamente. Compare os resultados obtidos com os calculados na etapa anterior.

11. Análise de variância com duas variáveis

Quando se ajustam os parâmetros a e b de uma reta $y = a + bx$ a um conjunto de dados experimentais $\{(x_i, y_i, s_i)\}, i = 1..N$, é possível identificar que fração da variação da grandeza é explicada pela relação funcional. No entanto, o mais frequente é que s_i seja desconhecido, mas igual para todos os pontos e possa ser estimado dos dados. Nesse caso, o próprio estimador da fração da variância atribuída à flutuação estatística tem uma flutuação estatística, o que complica o problema.

Etapa 1. Desenvolva a álgebra da regressão linear no caso $s_i = s$, igual para todos os pontos (veja seção 4.7 do texto). Calcule algebricamente o coeficiente de correlação *r* e explique por que ele dá a fração da variação de *y* explicada pela regressão (veja seção 8.9 do livro).

Etapa 2. Escolha valores numéricos para *a*, *b* e *s*, bem como um conjunto de coordenadas $\{x_i\}$ para simular o processo, usando a f.d.p. normal. Simule muitas vezes e faça um histograma dos valores obtidos de *s* e de *r*.

Etapa 3. Repita a etapa 2 para diversos valores de *b* e *s* e avalie os tamanhos de amostras necessárias para estimar os valores dessas grandezas com 10% de precisão relativa.

12. O método de Monte Carlo

Este seminário deve apresentar o princípio do método de Monte Carlo e ilustrá-lo com algumas aplicações. Monte Carlo um método de cálculo de integrais multidimensionais que se fundamenta na teoria da probabilidade, assim tem pouco conteúdo estatístico intrínseco a ele. No entanto, seu uso está em expansão e é um excelente método para avaliar o comportamento estatístico de variáveis aleatórias e processos sujeitos a erros, de modo que convém entendê-lo bem.

Etapa 1. Leia o livro "The Monte Carlo Method", por I.M. Sobol, pulando as seções 2.2 e 2.3 e o capítulo 3. São cerca de 50 páginas.

Etapa 2. Prepare o seminário em 5 partes:

i. Introdução ao método e descrição geral. Dê atenção particular ao método da inversão.
ii. Exemplo do sistema de suprimento
iii. Exemplo do transporte de nêutrons
iv. Exemplo do cálculo da integral definida
v. Tópicos avançados: amostragem de uma direção arbitrária e redução de variância

13. Resolução das equações do MMQ por um método gráfico

Considere que a relação entre a grandeza observada F e o tempo t, suposto conhecido exatamente, é

F(t) = N0*exp(-p*t) +a

Em que N0, p e a são constantes e a grandeza F tem distribuição normal.

A tabela abaixo dá um exemplo dos resultados experimentais obtidos

(inclua um exemplo dos seus dados!)

Um método gráfico para determinação dos parâmetros é o seguinte.
i. Fixe um valor de p, dentro da faixa de valores esperado para esse parâmetro.
ii. Com esse valor fixo, determine os parâmetros Ñ e ã que correspondem à estimativa do MMQ para os dados da tabela.

iii. Calcule

$$Q(p; \tilde{N}, \tilde{a}) = \sum_{i=1}^{n} \frac{(f_i - N \exp(-pt_i))^2}{(\sigma_i)^2}$$

Iv. Faça um gráfico de $Q(p; \tilde{N}, \tilde{a})$ em função de p para toda a faixa de possíveis valores dessa grandeza, repetindo os itens ii e iii para cada novo valor de p escolhido.

iv. Analise o gráfico: se existe um único mínimo na região e o comportamento da função Q é parabólico na vizinhança do mínimo até um valor igual ao mínimo + 25, os parâmetros ajustados são p_{minimo} e os valores de Ñ e ã obtidos com esse p; qui-quadrado é o valor de Q(p_{minimo} ; Ñ e ã). Determine a matriz de covariância dos parâmetros procedendo como no ajuste de parâmetros não-lineares, usando como estimativas dos parâmetros os valores que minimizam Q.

v. Se há vários mínimos, mas entre os dois menores valores a diferença é maior que 25, proceda como em iv – o menor mínimo serve como mínimo único. Se a região em torno do mínimo não for parabólica, converse com o professor.

vi. Se há vários mínimos, com diferenças menores que 9, você não tem estatística suficiente para extrair um resultado a respeito de p a partir desses dados. Tome mais dados, mude o processo, escolha outros tempos, procure outra amostra, etc.

vii. Se a distância entre os mínimos é maior que 9 e menor que 25, procure o professor!

Bibliografia

[Arfken] Mathematical Methods for Physicists, G.Arfken & H.Weber, Academic Press, 4ª edição (1995)

[Bevington] Data Reduction and Error Analysis for the Physical Sciences, P.Bevington, McGraw-Hill, 1969

[Conover] Practical Nonparametric Statistics, W.J.Conover, John Wiley & Sons Inc. 1971

[Eadie] Statistical Methods for Physicists, W.T.Eadie et al., North Holland Pub.Co. 1971

[Feller] An Introduction to Probability Theory and its Applications, John Wiley, 2ª Ed. (1957)

[Helene] Tratamento Estatístico de dados em FísicaExperimental, O.Helene, V.Vanin, Ed. Edgard Blücher, 2ª Ed., 1991

[James] A review of pseudorandom number generators, F.James, Computer Physics Communications 60(1990)329-344

[Kendall] The Advanced Theory of Statistics, M.Kendall, A.Stuart & J.K.Ord, Charles Griffin & Company Limited, London

[Mannhart] A Small Guide to Generating Covariances of Experimental Data, Report PTB–FMRD 84, Berlin, 1981. ISSN 0341–6666

[Marquardt] An Algorithm for Least-Squares Estimation of Nonlinear Parameters, D. Marquardt, SIAM J. Appl. Math. 11, 431-441, 1963

[Noether] Introdução à Estatística – Uma abordagem não paramétrica, G.E.Noether, Guanabara Dois, 1983

[Vuolo] Fundamentos da Teoria de Erros, J.H.Vuolo, Ed. Edgard Blu¨cher, 1992

[Youden] Statistical Methods for Chemists, W.J.Youden, John Wiley 1951

[Zar] J.H. Zar, Appl. Statist. 27(1978)n.3, 280-290