

Análise de erros simples: aleatórios ou sistemáticos?

- A média é uma estatística muito comum; é o que fazemos o tempo todo, por exemplo ao 'fazer uma integração' de um objeto fraco (i.e. efetuando uma exposição com tempo finito com um telescópio+instrumento). A variância da média é:

$$S_m^2 = E \left[\left(\frac{1}{N} \sum_{i=1}^N X_i - \mu \right)^2 \right]$$

- o que, após alguma manipulação é igual a:

$$S_m^2 = \frac{\sigma^2}{N} + \frac{1}{N^2} \sum_{i \neq j} E[(X_i - \mu)(X_j - \mu)]$$

- Ignorando o segundo termo por agora, o primeiro expressa uma crença bem estabelecida, de que o erro de uma média é inversamente proporcional a $\sqrt{N}$.
- Agora, sobre o segundo termo, distribuições com variância infinita à parte, o familiar resultado $\sqrt{N}$ só se mantém se esse termo for zero. Esse termo contém a *covariância* que é definida como:

$$\text{cov}[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)]$$

e está fortemente relacionado com o coeficiente de correlação entre x_j e x_i que veremos mais adiante.

Análise de erros simples: aleatórios ou sistemáticos?

- Quando os dados são todos independentes entre si (ou seja, não há correlação entre o ruído de um dado e do próximo) o termo da covariância desaparece. Nesse caso os erros são chamados de puramente *aleatórios* e decrescem com $\sqrt{N}$.
- Frequentemente esse não é o caso. Tome como exemplo uma série temporal de medida do fluxo de uma estrela qualquer. Num dado momento um nuvem passa em frente à estrela e as contagens diminuem. É muito provável que os próximos dados sofram o mesmo problema.

Análise de erros simples: aleatórios ou sistemáticos?

- Nesse caso a distribuição probabilística se torna multivariada e fatorizações simples não podem ser feitas:

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \int \int (x_i - \mu_i)(x_j - \mu_j)g(x_i, x_j|\dots)dx_i dx_j$$

de modo que é necessário entender melhor os erros observacionais (ou seja escrever a distribuição conjunta $g(x_i, x_j|\dots)$)

- Nesses casos mais complexos o erro da média converge muito mais lentamente do que $\sqrt{N}$
- A distinção que fazemos entre erros aleatórios e sistemáticos em termos estatísticos são um contínuo, com a covariância sendo zero em um dos extremos (erros aleatórios).

Análise de erros simples: propagação de erros

- Frequentemente o que queremos saber é uma função dos dados coletados. Conhecendo os erros nos dados como estimamos o erro na quantidade desejada?
- Se os erros são pequenos, o modo mais fácil é através da expansão de Taylor. Supondo as variáveis x, y, z , com erros independentes $\delta X, \delta Y, \delta Z$ e estamos interessados nas funções $f(x, y, z, \dots)$.
- A mudança em f causada pelo erro é, em primeira ordem:

$$\delta f = \left. \frac{\partial f}{\partial x} \right|_{x=X} \delta X + \left. \frac{\partial f}{\partial y} \right|_{y=Y} \delta Y + \left. \frac{\partial f}{\partial z} \right|_{z=Z} \delta Z + \dots$$

- Como a variância de uma soma é a soma das variâncias (e a variância é a expectativa do erro ao quadrado):

$$\text{var}[f] = \left(\left. \frac{\partial f}{\partial x} \right|_{x=X} \right)^2 \sigma_x^2 + \left(\left. \frac{\partial f}{\partial y} \right|_{y=Y} \right)^2 \sigma_y^2 + \left(\left. \frac{\partial f}{\partial z} \right|_{z=Z} \right)^2 \sigma_z^2 + \dots$$

Análise de erros simples: propagação de erros

- Essas considerações levam a um resultado muito conhecido para a combinação de medidas: dadas n estimativas independentes X_j , cada qual tendo erros σ_j , a melhor estimativa combinada é a seguinte média ponderada:

$$\bar{X}_w = \sum_{j=1}^n \bar{X}_j / \sigma_j^2 / \sum_{j=1}^n 1 / \sigma_j^2$$

- A variância de $\bar{X}_w$ é:

$$\sigma_w^2 = 1 / \sum_{j=1}^n 1 / \sigma_j^2$$

Exemplo W&J P. 63

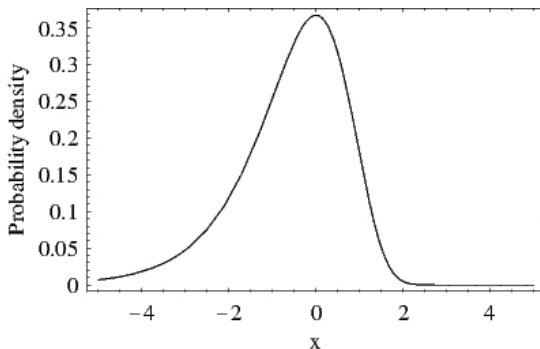
Análise de erros simples: combinando distribuições

- Frequentemente esse método não é bom o suficiente. Precisamos, provavelmente saber detalhes a distribuição probabilística da quantidade derivada. O caso mais simples é o de transformação da quantidade medida x , com distribuição g para uma quantidade derivada $f(x)$, cuja distribuição probabilística é dada por h .
- Como a probabilidade se conserva e lembrando que essas são distribuições são densidades de probabilidade temos:

$$h(f) df = g(x) dx$$

de modo que h envolve a derivada df/dx . Cuidados devem ser tomados se f não for monotônica.

Exemplo W&J P. 64



- Essa técnica rapidamente se torna difícil de aplicar para mais de uma variável. Seguem alguns resultados úteis

Análise de erros simples: combinando distribuições

- ❶ Suponha que temos duas variáveis x e y , identicamente distribuídas seguindo a função g . Qual a distribuição da sua soma: $z = x + y$? Para cada x devemos somar as probabilidades de todos os números $y = z - x$, que resultam no z de interesse. A distribuição probabilística $h(z)$ é então:

$$h(z) = \int g(z - x)g(x) dx,$$

onde as probabilidades foram simplesmente multiplicadas pela hipótese da independência.

A função h é chamada de função de autocorrelação de g .

Esse resultado é generalizável para a soma de várias variáveis.

- ❷ Frequentemente precisamos da distribuição do produto ou razão de duas variáveis. Sem maiores detalhes o resultado é o que se segue. Para $z = xy$, a distribuição de z é:

$$h(z) = \int \frac{1}{|x|} g(x)g(z/x) dx$$

e de $z = x/y$ é:

$$h(z) = \int |x| g(x)g(zx) dx.$$

Em quase todos os caso de interesse, essas integrais são muito difíceis para ser calculadas analiticamente.

Análise de erros simples: algumas estatísticas úteis e suas distribuições

- Para N dados X_i algumas estatísticas úteis são a média a variância amostral (S^2) e estatísticas de ordem. Se os X_i são independentes e variáveis gaussianas identicamente distribuídas, onde a gaussiana tem média μ e variância σ^2 , então:

- 1 O valor médio $\bar{X}$ segue uma distribuição gaussiana ao redor de μ , e com variância σ^2/N (TVC!)
- 2 A variância amostral σ_s^2 é distribuída como $\sigma^2\chi^2/(N-1)$, onde a variável “qui-quadrado” tem $N-1$ graus de liberdade.
- 3 A razão

$$\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma_s^2}$$

é distribuída como uma função t de Student, com $N-1$ graus de liberdade. Essa razão é de grande utilidade pois estima o quão longe da média verdadeira está a média dos dados.

- 4 Se nós temos duas amostras independentes de tamanhos N e M , geradas a partir da mesma distribuição gaussiana, então a razão da variância amostral σ_{s1}^2 e σ_{s2}^2 segue uma distribuição F (veremos mais sobre esses pontos no capítulo 5).

Análise de erros simples: algumas estatísticas úteis e suas distribuições

- **Estatísticas de ordem** são simplesmente os resultados do rearranjo dos dados X_i , em ordem crescente de valor, renomeados como $Y_1, Y_2, \dots$. Valores máximos e mínimos são de interesse e a mediana $Y_{N/2}$ (N ímpar) é um indicador robusto de localização.
- Podemos também formar estimativas robustas de espalhamento usando estatística de ordem, por exemplo o intervalo que contém 50% dos dados. Por isso ambas distribuições, de densidade e cumulativa são de interesse.
- suponha que a distribuição de x é $f(x)$, como distribuição cumulativa $F(x)$. Assim, a distribuição g_n da estatística de i -ésima ordem (Y_i) é:

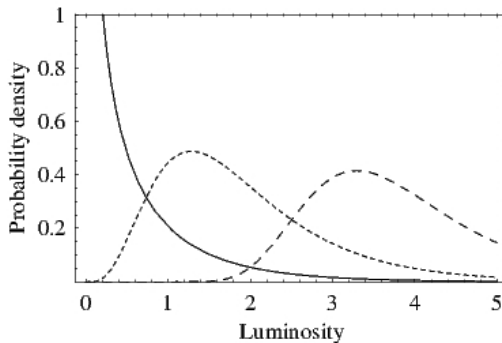
$$g_i(y) = \frac{N!}{(i-1)!(N-i)!} [F(y)]^{i-1} [1 - F(y)]^{N-i} f(y)$$

e sua distribuição cumulativa é

$$G_i(y) = \sum_{j=i}^N \binom{N}{j} [F(y)]^j [1 - F(y)]^{N-j}$$

Análise de erros simples: algumas estatísticas úteis e suas distribuições

Exemplo W&J P. 67



• Fim da aula 4