

Tratamento Estatístico de Dados em Física Experimental - Atividade 11

Geração de dados correlacionados

Faça as questões abaixo e depois transcreva suas respostas no formulário correspondente <https://forms.gle/YWycrsqEHErBA1JG8>. Assim como a Atividade 10, esta atividade deve ser enviada até as 23h59 do dia 03/novembro (quarta-feira). Até o final do prazo é possível revisar as respostas.

Considere o caso em que se deseja gerar pares de dados, a e b , gaussianos com valores verdadeiros a_0 e b_0 , desvios-padrões σ_a e σ_b e coeficiente de correlação ρ (ou seja, com covariância, $cov(a, b) = \rho\sigma_a\sigma_b$). Esses dados podem ser gerados a partir de dois números aleatórios gaussianos descorrelacionados, com valores verdadeiros nulos e desvios-padrão unitários, r_1 e r_2 , da seguinte forma:

$$\begin{aligned}a &= a_0 + \sigma_a \cdot r_1 \\b &= b_0 + \sigma_b \cdot (\rho \cdot r_1 + \sqrt{1 - \rho^2} \cdot r_2)\end{aligned}$$

Nota 1: Um procedimento mais geral, que permite gerar um vetor com diversos dados correlacionados (dois ou mais) com matriz de covariância desejada é apresentado no Anexo C.5 (pág. 62) do Suplemento 1 do GUM, disponível em https://www.bipm.org/utis/common/documents/jcqm/JCGM_101_2008_E.pdf

Nota 2: No Python, uma implementação do procedimento geral para o caso de dados gaussianos pode ser obtido com a função `random.multivariate_normal` do pacote numpy. Ver documentação em https://numpy.org/doc/stable/reference/random/generated/numpy.random.multivariate_normal.html

Vamos considerar um caso em que os valores verdadeiros de a e b sejam, respectivamente, $a_0 = 30$ e $b_0 = 20$ e que os desvios-padrões verdadeiros sejam iguais: $\sigma_a = \sigma_b = 2$, de modo que as simulações a seguir só irão diferir pelo coeficiente de correlação, ρ .

a) Gere $N = 500$ pares de dados com $\rho = +0.75$ e os represente em um gráfico de dispersão. Avalie qualitativamente se é fácil perceber alguma relação entre os valores de b de acordo com os valores de a .

a.1) Conte o número de vezes, n , em que os erros de a e b tem o mesmo sinal (o erro é a diferença entre o valor gerado e o valor verdadeiro). Usando seus conhecimentos sobre binomial, estime a incerteza de n .

a.2) Usando o valor de n calculado no item anterior com sua respectiva incerteza, determine a frequência relativa com que os erros de a e b têm o mesmo sinal, $f = \frac{n}{N}$, com sua respectiva incerteza.

Note que se os dados fossem estatisticamente independentes, a probabilidade dos erros terem o mesmo sinal seria 0,5.

a.3) Calcule a covariância amostral, V_{ab} , e a correspondente correlação amostral, R , dos N pares de valores de a e b gerados. Use as expressões fornecidas na Nota 2 para estimar as incertezas de V_{ab} e R . Compare com os valores de $cov(a, b)$ e ρ usados para gerar os dados.

Nota 1: a covariância amostral é calculada por $V_{ab} = \frac{1}{N-1} \sum_{i=1}^N (a_i - a_m)(b_i - b_m)$, onde a_m e b_m são os valores médios e a correspondente correlação amostral por $R = \frac{V_{ab}}{s_a s_b}$, onde s_a e s_b são os desvios-padrões amostrais.

Nota 2: A incerteza da covariância amostral pode ser estimada por $inc_V \cong s_a s_b \sqrt{\frac{(1+R^2)}{N-1}}$ e a do coeficiente de correlação por $inc_R \cong \frac{1-R^2}{\sqrt{N-1}}$

a.4) Para cada um dos N pares de valores de \mathbf{a} e \mathbf{b} , calcule a soma correspondente, $w = a + b$. Determine o desvio-padrão amostral de \mathbf{w} (a incerteza de cada valor de \mathbf{w}).

Nota: A incerteza do desvio-padrão amostral pode ser estimada por $inc_s \cong \frac{s}{\sqrt{2(N-1)}}$

a.5) Repita o item anterior para o caso da diferença entre \mathbf{a} e \mathbf{b} , $z = a - b$.

b) Repita o item **(a)** e seus subitens para o caso em que $\rho = -0,90$.

c) Faça apenas a geração dos dados com o gráfico de dispersão correspondente (isto é, o item **(a)** sem os subitens) para os casos $\rho = 0$, $\rho = +0,25$, $\rho = -0,5$ e $\rho = +0,95$.