**Introdução à Análise de dados em FAE** **(DATA)**

# Introdução a estatística para análise de dados em HEP

*Professores:* Dilson de Jesus Damião, Eliza Melo da Costa e Maurcio Thiel
*Name:* Bruno Kron Guandalini e Pedro Oliveira

## EXERCICIO 1

Deduza as equações para o ajuste linear para o caso em que as incertezas em y são diferentes para cada ponto.

(Sugestão: seção F.2 do livro Estimativas e Erros em Experimentos de Física)

RESPOSTA:

De acordo com o livro Estimativas e Erros em Experimentos de Física, para ajustar a reta $y(x) = ax + b$ a um conjunto de N pares de dados $(xi,\ yi)$, onde as incertezas entre os pares são distintas, os resíduos são ponderados pela seguinte expressão:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{[y_i - (ax_i + b)]^2}{\sigma_i^2} \tag{1}$$

Ao expandir a expressão em quadrados, temos:

$$\chi^2(a,b) = \sum_{i=1}^{N} \left( \frac{y_i^2}{\sigma_i^2} - \frac{2ax_iy_i}{\sigma_i^2} - \frac{2by_i}{\sigma_i^2} + \frac{a^2x_i^2}{\sigma_i^2} + \frac{2abx_i}{\sigma_i^2} + \frac{b^2}{\sigma_i^2} \right) \tag{2}$$

Depois, ao reescrever em termos das médias ponderadas, temos:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - 2a\sum_{i=1}^{N} \frac{x_iy_i}{\sigma_i^2} - 2b\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} + a^2\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} + 2ab\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + b^2\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{3}$$

Para encontrar os valores de $a$ e $b$ que minimizam $\chi^2$, derivamos parcialmente em relação a $a$ e $b$ e igualamos a zero:

$$\frac{\partial \chi^2}{\partial a} = -2\sum_{i=1}^{N} \frac{x_i(y_i - ax_i - b)}{\sigma_i^2} = 0 \tag{4}$$

$$\frac{\partial \chi^2}{\partial b} = -2\sum_{i=1}^{N} \frac{(y_i - ax_i - b)}{\sigma_i^2} = 0 \tag{5}$$

Isso nos leva ao sistema de equações:

$$\sum_{i=1}^{N} \frac{x_iy_i}{\sigma_i^2} - a\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - b\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = 0 \tag{6}$$

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - a\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} - b\sum_{i=1}^{N} \frac{1}{\sigma_i^2} = 0 \tag{7}$$

Resolvendo o sistema, encontramos:

$$a = \frac{\sum_{i=1}^{N} \frac{x_iy_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}\right)^2} \tag{8}$$

$$b = \frac{\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - a \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{9}$$

## EXERCICIO 2

A seção de choque de um processo definido como sinal pode ser obtida experimentalmente pela equação:

$$\sigma = \frac{N_{\text{Total}} - N_{\text{background}}}{L} \tag{10}$$

onde $N_{\text{Total}}$ é o número de eventos observados; $N_{\text{background}}$ é o número de eventos de fundo esperados e $L$ é a luminosidade integrada. Considerando para este caso que o número total de eventos observados foi 2567, o número de eventos de fundo esperado é 1223.5, e a luminosidade integrada é 25 fb$^{-1}$, com uma incerteza sistemática de 10%. Calcule o valor da seção de choque propagando separadamente as incertezas estatísticas (lembrando que aqui se trata de distribuição de Poisson) e sistemáticas.

RESPOSTA:
O exercício informa que a seção de choque é calculada pela equação:

$$\sigma = \frac{N_{\text{total}} - N_{\text{background}}}{L} \tag{11}$$

Onde $N_{\text{total}} = 2567$, $N_{\text{background}} = 1223.5$ e $L = 25\,\text{fb}^{-1}$. Calculando a seção de choque, temos:

$$\sigma = \frac{2567 - 1223.5}{25} = 53.74\,\text{fb} \tag{12}$$

Agora podemos calcular o erro estatístico do sistema. Por se tratar de uma distribuição de Poisson, temos que o erro é dado por:

$$\sigma^2 = \left(\frac{\sqrt{N_{\text{total}}}}{L}\right)^2 + \left(\frac{\sqrt{N_{\text{background}}}}{L}\right)^2 \tag{13}$$

$$\sigma_{\text{estatístico}} = \sqrt{\left(\frac{\sqrt{2567}}{25}\right)^2 + \left(\frac{\sqrt{1223.5}}{25}\right)^2} = 2.46\,\text{fb} \tag{14}$$

O exercício informa que o erro sistemático é de 10%, então podemos demonstrá-lo como:

$$\sigma_{\text{sistemático}} = \sigma \times 10\% = 5.37\,\text{fb} \tag{15}$$

Concluímos que a seção de choque do sinal com suas incertezas é:

$$\sigma = (53.74 \pm 2.46 \pm 5.37)\,\text{fb} \tag{16}$$

## EXERCICIO 3

A seleção de eventos em uma análise em Física de Altas Energias por busca de acoplamentos anômalos previu um número de eventos de fundo de 0.07 eventos após todos os cortes. Ao olhar para os dados, também após todos os cortes, se observou zero eventos. Diferentes modelos para acoplamentos anômalos previram diferentes números de eventos a serem observados, desde 0.09 até 35 eventos. Considerando que a contagem de eventos se dá através de uma pdf de Poisson, calcule até quantos eventos esperados podemos excluir com essa análise com 95% de C.L. (ver a tese "Study of WW Central Exclusive Production in the semileptonic channel with tagged protons at CMS detector" cap. 4.4)

RESPOSTA:
No exercício, a seleção de eventos para a busca de acoplamentos anômalos previu um número de eventos de fundo de $b = 0.07$ após todos os cortes. Diferentes modelos de acoplamentos anômalos previram uma variedade de números de eventos a serem observados, variando entre 0.09 e 35 eventos. Considerando que a contagem de eventos se dá através de uma PDF de Poisson, a probabilidade de observar $n$ eventos é dada pela expressão:

$$P(n; s, b) = \frac{(s+b)^n e^{-(s+b)}}{n!} \tag{17}$$

No caso específico em que $n = 0$, a probabilidade torna-se:

$$P(0; s, 0.07) = e^{-(s+0.07)} \tag{18}$$

Para excluir um modelo com um nível de confiança de 95%, é necessário que a probabilidade de observar zero eventos seja inferior a 5%, logo:

$$P(0; s, 0.07) < 0.05 \tag{19}$$

Substituindo na expressão, temos:

$$e^{-(s+0.07)} < 0.05 \tag{20}$$

Aplicando o logaritmo natural em ambos os lados, obtemos:

$$-(s + 0.07) < \ln(0.05) \tag{21}$$

Ajustando a expressão, teremos:

$$s + 0.07 > -\ln(0.05) \tag{22}$$

Calculando $-\ln(0.05)$:

$$-\ln(0.05) \approx 2.9957 \tag{23}$$

Sendo assim, a desigualdade será reescrita como:

$$s + 0.07 > 2.9957 \tag{24}$$

Isolando $s$, obtemos:

$$s > 2.9957 - 0.07 \implies s > 2.9257 \tag{25}$$

Concluímos que os parâmetros de acoplamentos anômalos que preveem um número esperado de eventos superior a $s > 2.93$ podem ser excluídos com um nível de confiança de 95%. Levando em consideração o intervalo de eventos previstos pelos modelos, que varia de 0.09 a 35 eventos, todos os modelos que preveem até 2.93 eventos permanecem como potenciais candidatos, enquanto aqueles que preveem um número de eventos maior que 2.93 são inviáveis de acordo com a análise realizada.

**EXERCICIO 4**

Mostre que a melhor função que se adequada aos dados é quando $\chi^2/ndf -> 1$. (ver Vuolo 12.6)

RESPOSTA:

A estatística $\chi^2$ é uma medida utilizada para avaliar a qualidade de um ajuste, definida por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{\sigma_i^2} \tag{26}$$

onde:

- $y_i$ é o valor observado no experimento.
- $f(x_i)$ é o valor previsto pela função de ajuste no ponto $x_i$.
- $\sigma_i$ representa a incerteza associada ao valor $y_i$.
- $n$ é o número total de dados.

Essa métrica $\chi^2$ fornece a soma das discrepâncias quadradas normalizadas pelas incertezas. Quanto melhor o ajuste, menor será o valor de $\chi^2$, uma vez que as discrepâncias serão menores.

O número de graus de liberdade ($ndf$) é calculado como:

$$ndf = n - p, \tag{27}$$

onde:

- $n$ é o número total de pontos de dados.
- $p$ é o número de parâmetros ajustados no modelo.

A quantidade $\chi^2/ndf$ é usada para avaliar a qualidade do ajuste de maneira mais robusta. Se as incertezas $\sigma_i$ forem estimadas corretamente e o modelo ajustado for adequado, espera-se que $\chi^2/ndf \approx 1$.

- Se $\chi^2/ndf > 1$: Um valor significativamente maior que 1 sugere que o modelo não descreve bem os dados. Isso pode ocorrer devido a:
  - Um modelo inadequado para os dados.
  - Subestimação das incertezas experimentais.
- Se $\chi^2/ndf < 1$: Um valor menor que 1 pode indicar que:
  - As incertezas foram superestimadas.
  - O modelo pode estar sobreajustando os dados (*overfitting*).

A condição $\chi^2/ndf \approx 1$ indica que a função ajustada é apropriada, ou seja, descreve os dados dentro das incertezas estatísticas esperadas. Isso significa que as discrepâncias observadas são consistentes com as flutuações devido às incertezas dos dados. Portanto, a melhor função para um ajuste ocorre quando $\chi^2/ndf \to 1$, garantindo que o modelo é compatível com os dados e que as incertezas foram corretamente estimadas.