Métodos Estatísticos em Física Experimental

Métodos Estatísticos em Física Experimental

Vitor Oguri

Departamento de Física Nuclear e Altas Energias
Instituto de Física Armando Dias Tavares
Universidade do Estado do Rio de Janeiro (UERJ)

1ª Edição

Direção editorial: José Roberto Marinho

Capa: Fabrício Ribeiro
Projeto gráfico e diagramação: Vitor Oguri

Edição revisada segundo o Novo Acordo Ortográfico da Língua Portuguesa

Dados Internacionais de Catalogação na publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Oguri, Vitor
Métodos estatísticos em física experimental / Vitor Oguri. – São Paulo: Editora Livraria da Física, 2017.

1. Dados - Análise 2. Física - Experiências 3. Física - Métodos estatísticos
4. Probabilidades I. Título.

Bibliografia
ISBN 978-85-7861-474-4

1. Física - Estudo e Ensino 2. Física - Pesquisa 3. Prática de ensino 4. Professores - Formação 5. Sociologia educacional.

17-03445 CDD-530

Índices para catálogo sistemático:
1. Experimentos: Física: Métodos estatísticos 530

LF EDITORIAL

Editora Livraria da Física
www.livrariadafisica.com.br

9 A Nobuo Oguri, in memoriam

10 Se deseja saber a essência do método científico, não preste atenção ao que lhe possa dizer um cientista. Observe o que ele faz.

Albert Einstein

Prefácio

Este livro é fruto da experiência do autor em análise de dados em experimentos de Física, desde sua formação inicial nos anos de 1980, na área de raios X e radiação síncrotron, nos laboratórios da faculdade de engenharia física da Universidade de Tóquio e na Photon Factory do KEK,[1] até a sua participação em grandes experimentos de Altas Energias; DZero do Fermilab[2] e CMS do CERN,[3] nos anos de 1991 até 2013.

O material selecionado é abordado no corpo do texto de modo intuitivo, esperando-se que o leitor tenha familiaridade apenas com os tópicos básicos de um curso de Cálculo Diferencial e Integral, e com alguns conceitos da Estatística, todos temas abordados regularmente nos ciclos básicos dos cursos de Física, Química e Engenharias.

O texto que serviu de ponto de partida do livro corresponde às notas de aula da disciplina de Tratamento estatístico de dados em Física, ministrada pelo autor em diversos períodos, ao longo dos últimos 10 anos, na graduação e no programa de pós-graduação do Instituto de Física Armando Dias Tavares da Universidade do Estado do Rio de Janeiro (Uerj).

Apesar de muitos exemplos serem tomados da área de Altas Energias, as ferramentas e os métodos estatísticos apresentados podem ser utilizados em qualquer experimento que envolva processos aleatórios.

O autor agradece aos colegas do Departamento de Física de Altas Energias da Uerj, Alberto Franco de Sá Santoro, Francisco Caruso, José Roberto Mahon e Dilson de Jesus Damião, pela leitura, críticas e sugestões ao texto.

Um agradecimento especial à Stella Maris Amadei pela cuidadosa leitura de diferentes versões de todos os capítulos e pelas sugestões de estilo.

V.O.

[1] Acelerador Síncrotron do Laboratório de Altas Energias do Japão, em Tsukuba.
[2] Fermi National Accelerator Laboratory, Batavia, Illinois, USA.
[3] Centro Europeu de Física de Partículas, Genebra, Suíça.

Sumário

5

Determinação de parâmetros

> Os problemas de estimação surgem quando conhecemos, ou consideramos que conhecemos, a forma da distribuição de frequências de uma população como uma função de um ou mais parâmetros desconhecidos e, a partir de uma amostra de dados, desejamos estimar os valores desses parâmetros.
>
> R. A. Fisher

5.1 Origens dos métodos estatísticos

As origens dos métodos e testes estatísticos modernos remontam aos trabalhos dos britânicos Francis Galton (1822-1911), Karl Pearson (1857-1936), William Gosset (1876-1937), Ronald Fisher (1890-1962) e do polonês Jerzy Neyman (1894-1981).

Em sua obra clássica, The grammar of science,[4] em 1892, Pearson estabelece os modelos estatísticos como alternativa à visão determinística do século XIX, com base nos seguintes princípios:

- todo experimento está sujeito a efeitos imprevistos e não observáveis;
- os resultados de um experimento obedecem a distribuições de probabilidades caracterizadas por parâmetros e propriedades como: valor esperado, variância, assimetria e curtose.

Para Pearson, as distribuições limites de probabilidades descreviam verdadeiramente a coleção de dados (medidas) resultante de um experimento e, a partir de um grande número de medições, os parâmetros da distribuição real das medidas ou dos dados dos experimentos poderiam ser determinados.

Por outro lado, Fisher, em seus trabalhos, sintetizados nos textos Statistical methods for research workers (1925) e The design of experiments (1935), [5] considera

[4] Reimpresso por Dover Publications, Inc., em 2004.
[5] Ambos reimpressos por Oxford University Press, em 2003.

que os dados constituem uma amostra aleatória de uma população hipotética e, a partir de um experimento, obtêm-se apenas os estimadores dos parâmetros da distribuição hipotética dos dados.

Ao contrário dos parâmetros (hipotéticos), os estimadores são aleatórios e devem ser avaliados, tanto segundo as distribuições limites de probabilidades, quanto aos seguintes critérios:

- consistência – quanto maior o número (N) de dados em uma amostra, mais próximo um estimador $\hat{a}$ deve estar do valor (a) do parâmetro;

$$\lim_{N\to\infty} \hat{a} = a$$

- eficiência – quanto menor a variância associada, mais eficiente é o estimador;

$$V(\hat{a}_1) < V(\hat{a}_2) \implies \hat{a}_1 \text{ mais eficiente}$$

- não-tendenciosidade – o valor esperado de um estimador $E(\hat{a})$ deve ser igual ao valor (a) do parâmetro.

$$E(\hat{a}) = a$$

Para Pearson e Fisher, ou para a escola clássica, os parâmetros (mesmo que sejam hipotéticos) são fixos e os estimadores, aleatórios; para a escola bayesiana, no entanto, tanto os parâmetros como os estimadores são aleatórios.

5.2 O método da máxima verossimilhança de Fisher

Seja $p(x|\theta)$ a distribuição de probabilidades para as medidas de uma grandeza x, em que θ é o parâmetro da distribuição. Para uma amostra $(x_1, x_2, \ldots, x_N)$ de N medidas independentes de x, a probabilidade associada a essa sequência particular de medidas é dada por

$$\prod_{i=1}^{N} p(x_i|\theta)$$

Se apenas a forma funcional da distribuição de probabilidades a priori for conhecida, isto é, o parâmetro for desconhecido, a função definida por

$$\boxed{\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = K \prod_{i=1}^{N} p(x_i|\theta)} \tag{5.1}$$

sendo K é uma constante arbitrária, denominada função de verossimilhança [2],[6] quantifica o quão verossímil é qualquer hipótese relativa ao valor do parâmetro.

Nesse sentido, para uma dada amostra de dados, se θ_A e θ_B representam dois possíveis estimadores para o parâmetro, e

$$\mathcal{L}(\theta_A) > \mathcal{L}(\theta_B)$$

diz-se que θ_A é um estimador mais verossímil para o parâmetro do que θ_B.

> Tomando-se o exemplo da seção ??: da caixa contendo 10 bolas, entre vermelhas e azuis, da qual é extraída uma amostra com reposição de 3 bolas vermelhas e uma azul.
>
> Para se estimar a quantidade x de bolas azuis contidas na caixa, a função de verossimilhança $\mathcal{L}(x)$ correspondente à amostra obtida será dada pela probabilidade binomial de que em 4 tentativas ($N = 4$) de extração de bolas azuis haja apenas um sucesso ($m = 1$).
>
> $$\mathcal{L}(x) \;\propto\; B(m = 1|N = 4; p = x/10)$$
>
> ou seja,
>
> $$\mathcal{L}(x) \;\propto\; \frac{x}{10}\left(1 - \frac{x}{10}\right)^3 = \frac{x(10-x)^3}{10000} \qquad \Rightarrow \qquad \mathcal{L}(x) = x(10-x)^3$$
>
> De acordo com a Tab. 5.1, representada na Fig. 5.1, que mostra os valores de $\mathcal{L}(x)$ para todos os possíveis valores de x, e segundo, ainda, o princípio de Fisher, o valor 3 é a melhor estimativa para a quantidade de bolas azuis.

Ao se obter três amostras de quatro tentativas de extração de bolas azuis, por exemplo, duas amostras de um sucesso ($m = 1$) e uma de dois sucessos ($m = 2$), a função de verossimilhança é dada por (Fig. 5.2)

$$\mathcal{L}(x) \;\propto\; B(1|4; x/10)^2 \times B(2|4; x/10)$$

[6] Likelihood function

Tabela 5.1: Função de verossimilhança para uma amostra de um sucesso ($m = 1$) em 4 tentativas ($N = 4$) de extração de bolas azuis de uma caixa contendo 10 bolas.

x	$\mathcal{L}(x)$
0	0
1	729
2	1024
3	1029
4	864
5	625
6	384
7	189
8	64
9	9
10	0

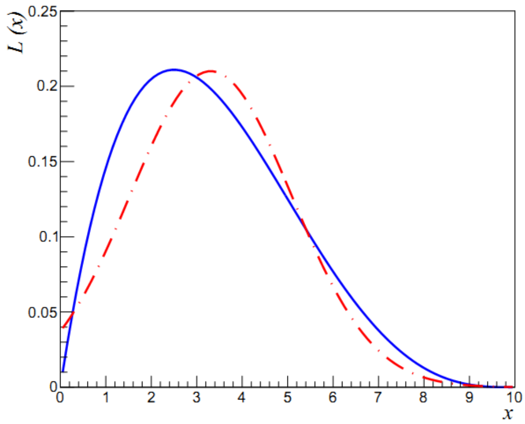

Figura 5.1: Função de verossimilhança (linha contínua) para uma amostra de um sucesso em 4 tentativas de extração de bolas azuis de uma caixa contendo 10 bolas. A linha pontilhada é uma distribuição gaussiana de mesmo valor esperado e desvio padrão da função de verossimilhança normalizada.

o que implica, igualmente, o valor 3 como a melhor estimativa para a quantidade de bolas azuis.

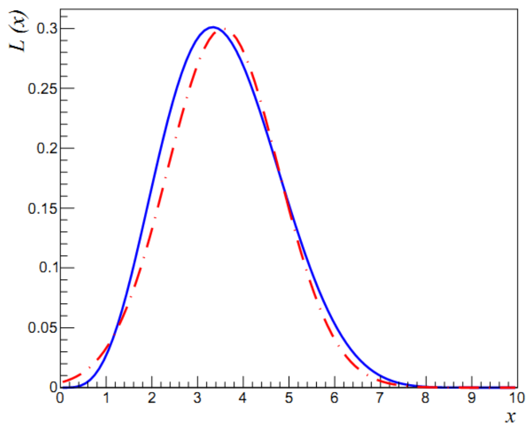

Figura 5.2: Função de verossimilhança (linha contínua) para 2 amostras de um sucesso, e uma de dois sucessos em 4 tentativas de extração de bolas azuis de uma caixa contendo 10 bolas. A linha pontilhada é uma gaussiana de mesmo valor esperado e desvio padrão.

Desse modo, Fisher considera que o estimador de máxima verossimilhança[7] ($\hat{\theta}$) para o parâmetro θ é aquele que maximiza a função de verossimilhança, ou seja,

$$\left.\frac{\partial \mathcal{L}}{\partial \theta}\right|_{\hat{\theta}} = 0 \tag{5.2}$$

Uma vez que o logaritmo de $\mathcal{L}$ atinge seu máximo para o mesmo valor de θ e $\mathcal{L}$, a condição de máxima verossimilhança é expressa como

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\hat{\theta}} = 0 \tag{5.3}$$

ou seja, o estimador de máxima verossimilhança é raiz da equação anterior (Eq. 5.3).

[7] Maximum likelihood estimator.

Mesmo que o domínio da variável aleatória x seja contínuo, como usualmente as medidas de uma grandeza, as amostras constituem sequências discretas de valores organizados em M classes de intervalos $\varDelta_i$, como

$$\Big\{(x_1, x_2 = x_1 + \varDelta_1); (x_2, x_3 = x_2 + \varDelta_2); \ldots ; (x_M, x_{M+1} = x_M + \varDelta_M)\Big\}$$

Nesse caso, a probabilidade associada a essa sequência particular é dada por

$$\int_{x_1}^{x_2} \int_{x_2}^{x_3} \ldots \int_{x_M}^{x_{M+1}} \rho(x_1|\theta)\rho(x_2|\theta)\ldots\rho(x_M|\theta)\ \mathrm{d}x_1\,\mathrm{d}x_2\ldots\,\mathrm{d}x_M$$

e define-se a função de verossimilhança por

$$\boxed{\mathcal{L}(x_1, x_2, x_3 \ldots x_M; \theta) = K\ \prod_{i=1}^{M} \rho(x_i|\theta)} \tag{5.4}$$

em que K é uma constante arbitrária.

5.3 Exercícios

5.3.1) Ao colidir com a superfície terrestre, um meteoro provoca uma cratera. A relação esperada entre o diâmetro (D) da cratera e a energia cinética (E) do meteoro no instante do impacto é dada por

$$D = kE^{1/4}$$

em que k é uma constante.

A tabela abaixo mostra os diâmetros das depressões causadas pelo impacto de diversas esferas de aço sobre a areia contida em uma caixa, e as correspondentes incertezas (ϵ_D) e energias cinéticas das esferas ao colidirem com a areia da caixa. As esferas são utilizadas para simularem a queda de meteoros.

E(J)	D(cm)	ϵ_D(cm)
0,07	4,9	0,3
0,18	6,7	0,3
0,30	7,3	0,4
0,45	8,1	0,4
0,69	9,2	0,4

A partir de um ajuste linear, determine uma estimativa para o expoente da relação esperada entre a energia e o diâmetro.

Referências Bibliográficas

[1] P. A. M. Dirac, The Principles of Quantum Mechanics, Oxford University Press, 1958.

[2] R. A. Fisher, On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society, v. 222, 309-368, 1922.

[3] R. A. Fisher, Applications of "Student's" Distribution, Metron 5, 90-104, 1925.

[4] S. M. Stigler, The History of Statistics: The Measurement of Uncertainty before 1900, Harvard University Press, 1986.

[5] Student, The Probable Error of a Mean, Biometrika, Vol. VI (1908), 1-25, 1900.