Tratamento de dados em Física

Métodos e testes estatísticos

V. Oguri

Departamento de Física Nuclear e Altas Energias (DFNAE)
Programa de Pós-graduação em Física (PPGF)
Instituto de Física Armando Dias Tavares (IFADT)
Universidade do Estado do Rio de Janeiro (UERJ)

Rio de Janeiro - Brasil
10 de agosto de 2011

Karl Pearson (1857-1936)

The Grammar of Science – 1892

- Modelos estatísticos como alternativas à visão determinística do séc. XIX
- Todo experimento está sujeito a efeitos imprevistos e não observáveis
- Os resultados $\{x_i\}$ de um experimento obedecem a certas distribuições estatísticas, $\rho(x)$, que são caracterizadas por alguns parâmetros: média, desvio-padrão, simetria e curtose
- Os observáveis na Ciência são as funções de distribuições estatísticas que descrevem as probabilidades associadas às observações
- Método dos momentos e teste de χ^2 (1900)

Ronald Fisher (1890-1962)

Statistical Methods for Research Workers – 1925
The Design of Experiments – 1935

- Todo experimento deve começar com um modelo matemático que estime os resultados esperados
- Os estimadores são aleatórios e devem ser avaliados segundo suas distribuições de probabilidades, e aos seguintes critérios:
  ▷ consistência – quanto maior o tamanho (N) de uma amostra mais próximo um estimador $\hat{a}$ estará do valor esperado $\langle a \rangle$ do parâmetro $\lim_{N\to\infty} \hat{a} = \langle a \rangle$
  ▷ eficiência – quanto menor a variância associada mais eficiente é o estimador $V(a_1) > V(a_2) \implies a_2$ mais eficiente
  ▷ não-tendenciosidade – o valor médio de um estimador deve tender ao valor esperado do parâmetro $\overline{\hat{a}} \to \langle a \rangle$
- Graus de liberdade (1922) e Likelihood (1925)

Pearson × Fisher

- **Pearson**
  As distribuições descrevem verdadeiramente os dados (medidas) resultantes de um experimento, no sentido de que a partir de um grande número de medições pode-se determinar os parâmetros da distribuição real das medidas.

- **Fisher**
  As medições resultam de uma população hipotética, e a partir de um experimento obtém-se apenas os estimadores dos parâmetros das distribuições hipotéticas dos dados.

Método da Máxima Verossimilhança de Fisher

- $\rho(x_i|\theta) \rightarrow$ distribuição de probabilidades (PDF) para as medidas de uma grandeza x, onde θ é um parâmetro da distribuição
- para uma amostra $A = \{x_1, x_2, x_3 \ldots x_N\}$ de N medidas independentes, a probabilidade associada é dada por

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

- se a PDF ρ e o parâmetro são corretos, espera-se que em relação a qualquer outra distribuição ou parâmetro, essa probabilidade seja máxima.

- A função definida por

$$\mathcal{L}(\theta|x_1, x_2, x_3 \ldots x_N) = K \prod_{i=1}^{N} \rho(x_i|\theta)$$

quantifica o quão verossímil é o valor do parâmetro para uma dada amostra de dados, no sentido que, se θ_A e θ_B são dois valores possíveis para o parâmetro e

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

diz-se que θ_A é uma estimativa mais verossímil para o parâmetro do que θ_B.

- Os estimadores de máxima verossimilhança (ML) são aqueles que maximizam a chamada **função de verossimilhança** (likelihood).
- Fisher mostrou que esses estimadores são “geralmente” consistente, os mais eficientes, e a tendenciosidade pode ser calculada e, portanto, subtraída.

Exemplo

De uma caixa que contém 10 bolas, entre vermelhas e azuis, é extraída uma amostra com reposição de 3 bolas vermelhas e uma azul. Quantas bolas azuis há na caixa?

Para se estimar a quantidade x de bolas azuis contidas na caixa, a função de verossimilhança $\mathcal{L}(x)$ correspondente a amostra obtida será dada pela probabilidade binomial de que em 4 tentativas ($N = 4$) de extração de bolas azuis houve apenas um sucesso ($m = 1$).

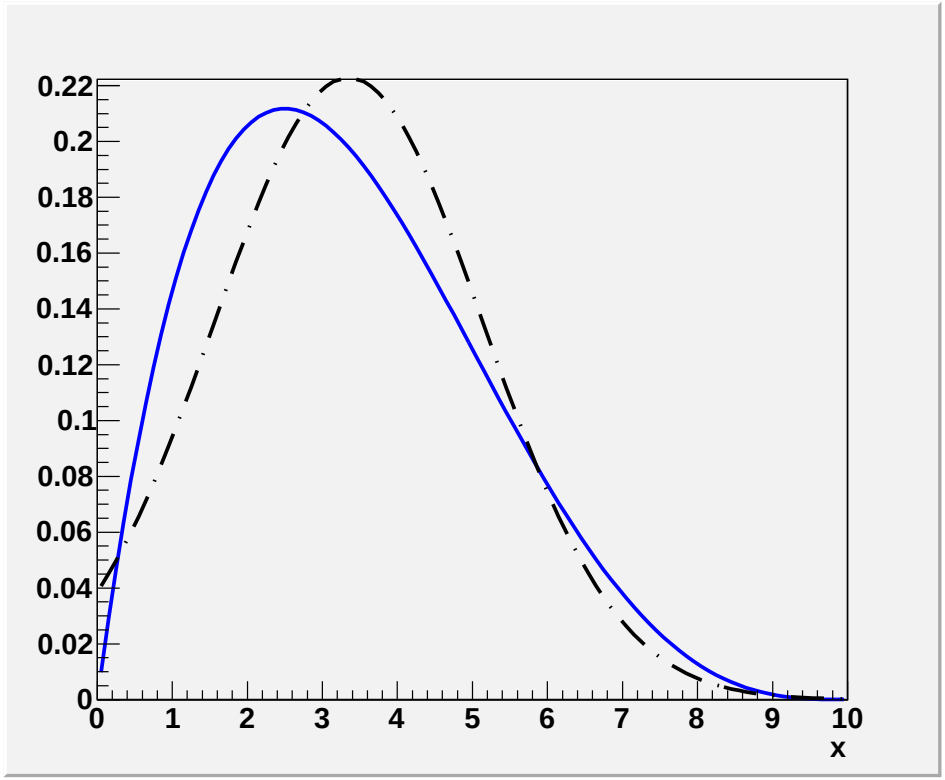
$$\mathcal{L}(x) \propto B(m = 1, N = 4, p = x/10)$$

$$\mathcal{L}(x) = 4\frac{x}{10}\left(1 - \frac{x}{10}\right)^3 = \frac{x(10 - x)^3}{2500}$$

valores de $\mathcal{L}(x)$ para todos os possíveis valores de x

x	$\mathcal{L}(x) \times 2500$
0	0
1	729
2	1024
3	1029
4	864
5	625
6	384
7	189
8	64
9	9
10	0

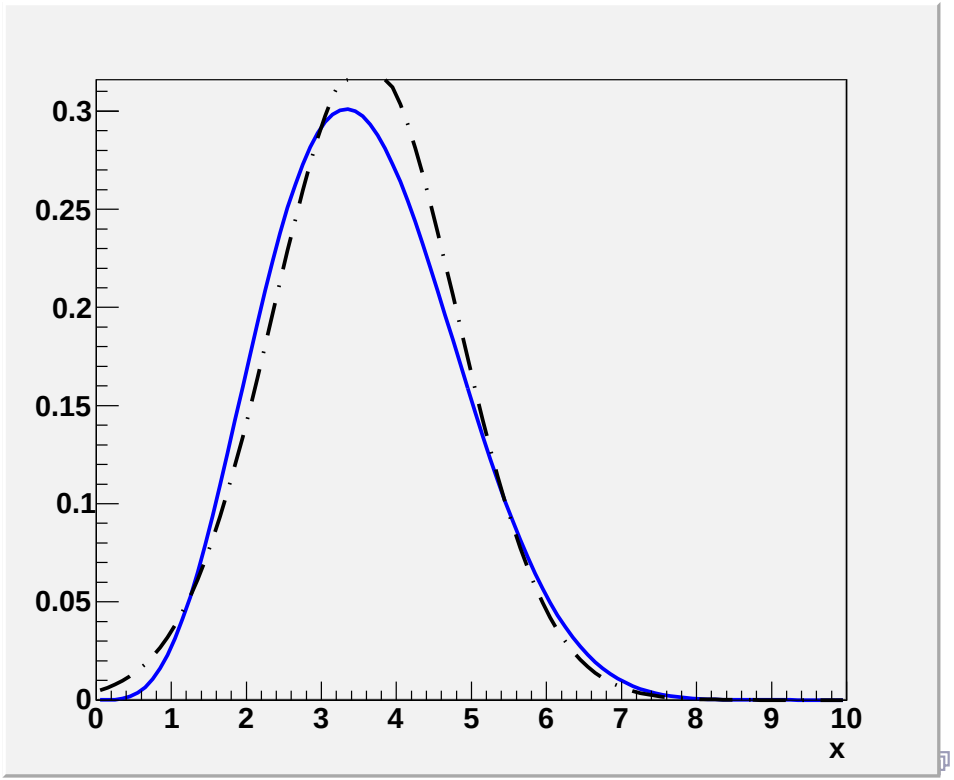
▷ Segundo o Princípio de Fischer, a melhor estimativa para a quantidade de bolas azuis é 3.



Comparação da função de verossimilhança normalizada (linha contínua – azul) com uma distribuição gaussiana (linha pontilhada – preta) de mesma média e variância.

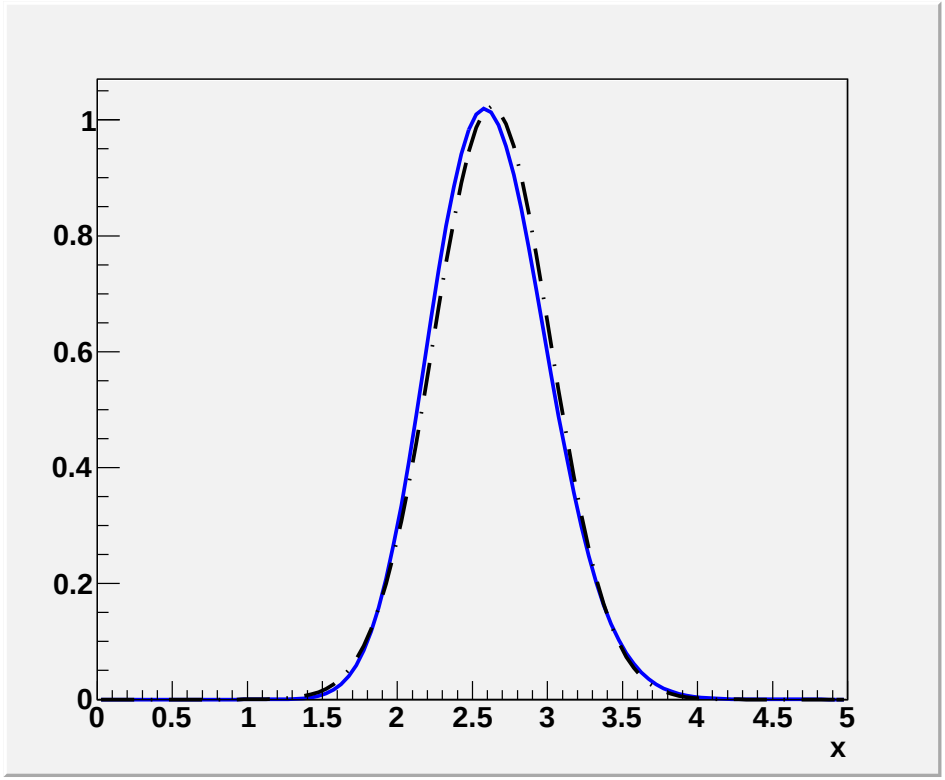
Aumentando-se o número de amostras de quatro tentativas, por exemplo, para duas amostras de um sucesso (uma bola azul) e uma de dois sucessos (duas bolas azuis), a função de verossimilhança é dada por

$$\mathcal{L}^3(x) = B(1,4,x/10)^2 \times B(2,4,x/10)$$



para 30 amostras de um sucesso (uma bola azul) e uma de dois sucessos (duas bolas azuis)

$$\mathcal{L}^{31}(x) = B(1,4,x/10)^{30} \times B(2,4,x/10)$$



Estimativas de parâmetros

Limite da função de verossimilhança

- espera-se que a função de verossimilhança, como função de um parâmetro θ, represente uma distribuição de "probabilidades" cujo máximo seja dado por um valor situado em torno do valor esperado para o parâmetro.
- para um número "suficientemente grande" de medidas, pode-se expandir o logaritmo da função de verossimilhança em torno de seu máximo,

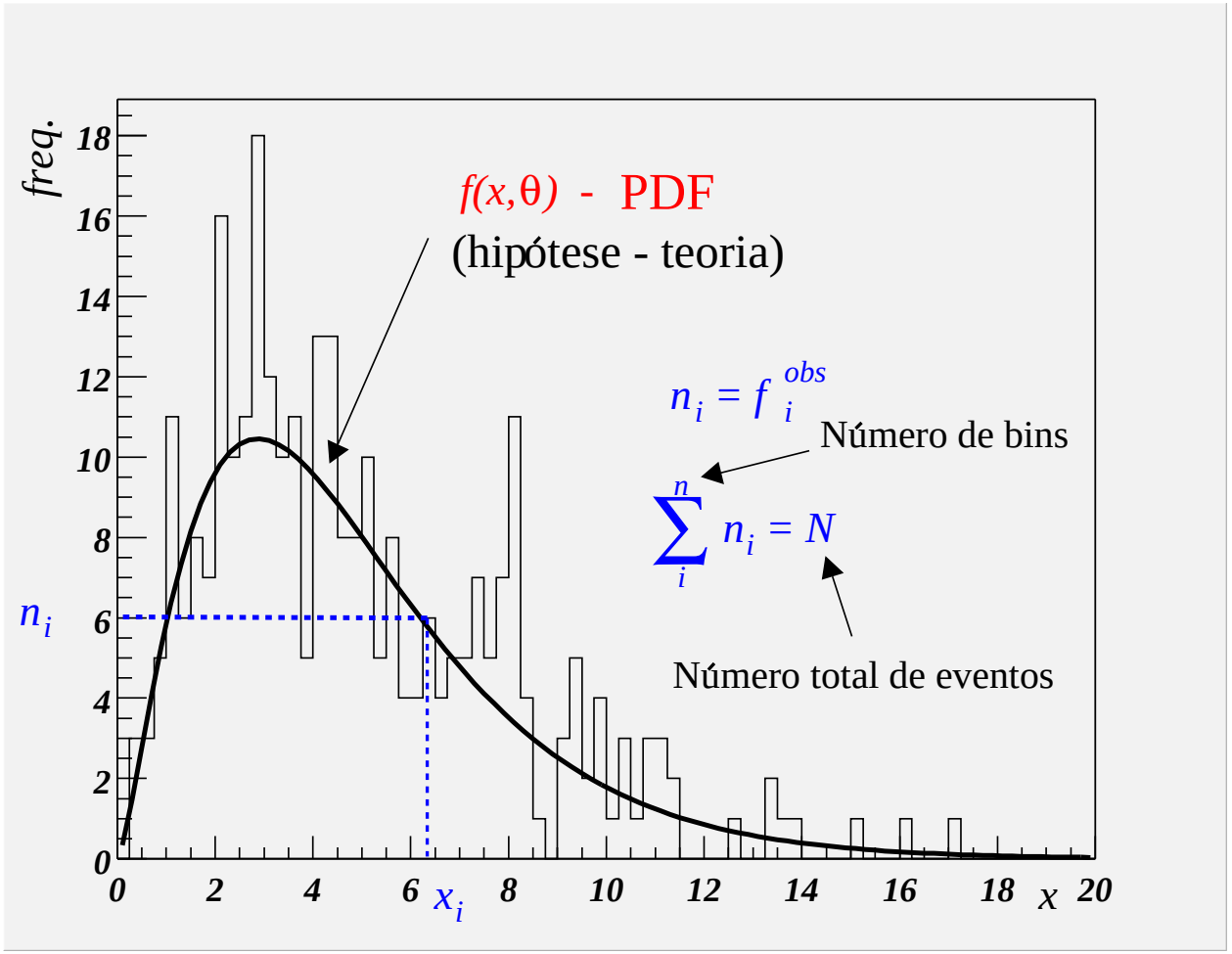
$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\hat{\theta}) - \frac{1}{2} \left| \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \right|_{\hat{\theta}} (\theta - \hat{\theta})^2$$

Limite da função de verossimilhança

para grandes amostras (teorema do limite central), a função de verossimilhança pode ser aproximada por uma **distribuição gaussiana**

$$\mathcal{L}(\theta) \propto e^{-\frac{1}{2}\left|\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\hat{\theta}}(\theta-\hat{\theta})^2}$$

Ajuste de funções a histogramas



Ajuste de funções e estimativa de parâmetros

$p_i(\theta) = \int_{x_i^{\min}}^{x_i^{\max}} f(x,\theta)dx$ (probabilidade associada a cada bin)

$f_i^{\text{obs}} = n_i$ (freqüência observada)

$f_i^{\text{esp}} = \epsilon_i(\theta) = Np_i(\theta)$ (freqüência esperada segundo PDF)

$\theta = (\theta_j) = (\theta_1, \theta_2, \ldots, \theta_p)$ (parâmetros da PDF)

- Como determinar os parâmetros?
- Quão bem a PDF $f(x,\theta)$ se ajusta aos dados?

Método dos mínimos quadrados de Lagrange

$$S = \sum_{i=1}^{n} \frac{[n_i - \epsilon_i(\theta)]^2}{\epsilon_i}$$

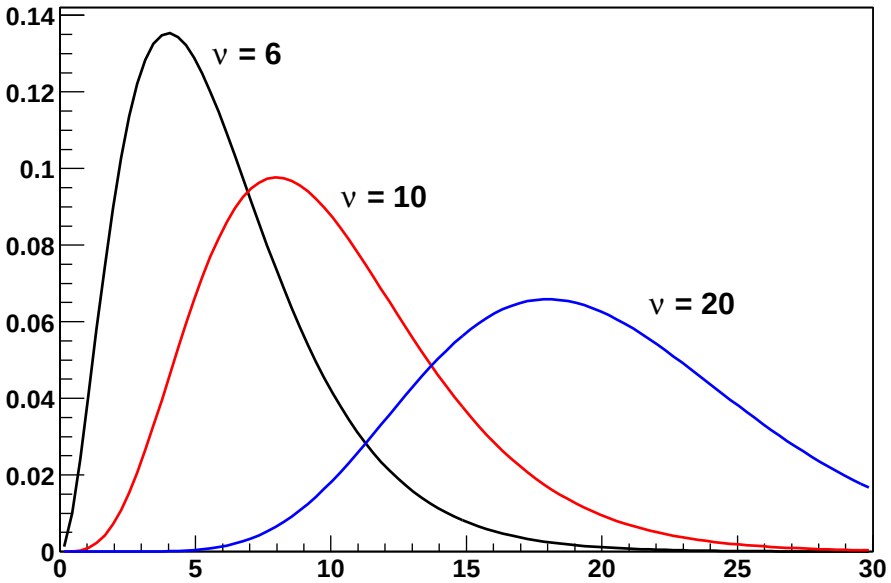
- parâmetros θ_j são aqueles que minimizam S

$$(p+1) \text{ equações} \quad \begin{cases} \dfrac{\partial S}{\partial \theta_j} = 0 \quad j = 1, \ldots, p \\ \text{de vínculos} \quad \displaystyle\sum_{i=1}^{n} n_i = N \end{cases}$$

- $S \to (n - p - 1)$ termos independentes
- $\nu = (n - p - 1)$ nº de graus de liberdade

Teste de χ^2

Pearson – *S* obedece à distribuição de χ^2



$$\rho(\chi^2|\nu) = A\left(\frac{\chi^2}{2}\right)^{\nu/2-1} e^{-\chi^2/2}$$

- $\mu = \nu$
- $\sigma^2 = 2\,\nu$
- $\rho_{\text{max}} \simeq \rho(\nu)$

Qualidade do ajuste

- parâmetros $(\theta_j) \Longrightarrow S = \chi^2_{\text{calc}}$
- $S = \chi^2_{\text{calc}} \simeq \nu \quad \Longleftrightarrow \quad \rho(\chi^2_{\text{calc}}) \simeq \rho_{\text{max}}$ O ajuste é bom
- se $p(\chi^2 > \chi^2_{\text{calc}}) = \int_{\chi^2_{\text{calc}}}^{\infty} \rho(\chi^2|\nu) d\chi^2$ for pequeno

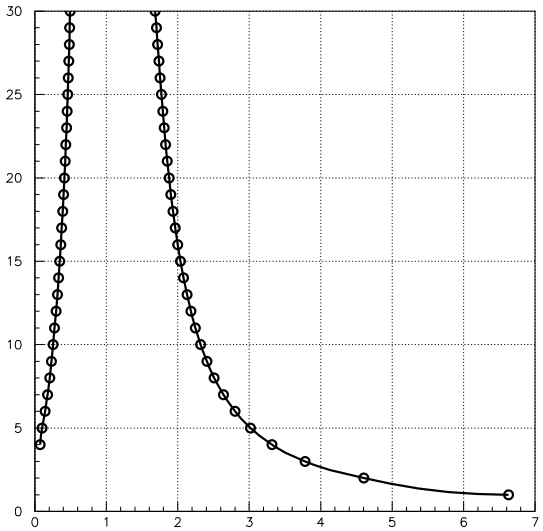
  $\Downarrow$

  χ^2_{calc}(grande) o ajuste não é bom.

- se $p(\chi^2 > \chi^2_{\text{calc}}) \simeq 1 (\chi^2_{\text{calc}} \to 0)$

  ajuste não compatível com os erros.

Intervalo de confiança



ordenada – ν (graus de liberdade - dgf)
abscissa – χ^2/ν

$$\int_{\chi^2_{\text{inf}}}^{\infty} \rho(\chi^2|\nu) d\chi^2 = 0.95 \qquad \int_{\chi^2_{\text{sup}}}^{\infty} \rho(\chi^2|\nu) d\chi^2 = 0.05$$

$$(\chi^2_{\text{sup}} - \chi^2_{\text{inf}}) \Longleftrightarrow \text{Intervalo de confiança de 90\%}$$

Exemplo

Lançamento de dados

ocorrência de uma face, valor ou evento (i) $\iff$ processo aleatório (probabilidade - p_i)

- a posteriori $\rightarrow p_i = \lim_{N\to\infty}\left(\frac{n_i}{N}\right)$
- a priori $\rightarrow p_i = \frac{1}{6}$
- $\sum_{i=1}^{n=6} p_i = 1$ (normalização)
- $\langle i\rangle = \sum_{i=1}^{n=6} i \times p_i = \frac{1}{6}\underbrace{(1+2+3+4+5+6)}_{21} = 3.5$ (média)

Amostra de N lançamentos

$N = 120$

i	1	2	3	4	5	6
n_i	16	19	27	17	23	18
ϵ_i	20	20	20	20	20	20

freq. observadas $(f_i^{\text{obs}} = n_i) \quad \rightarrow \{n_1, \ldots, n_6\}$

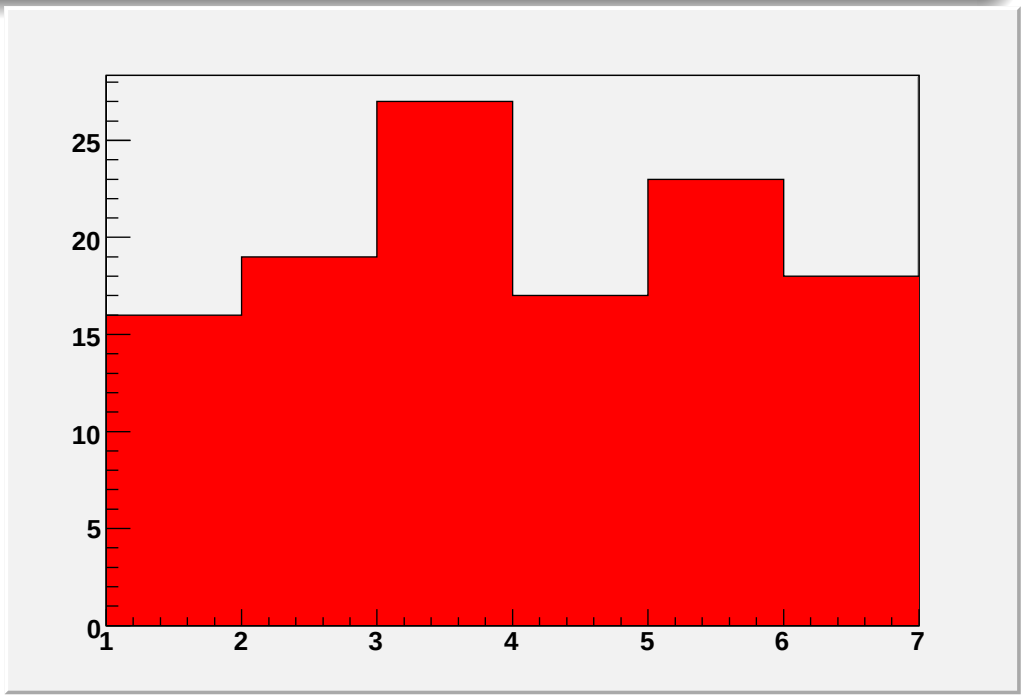
$$\sum_{i=1}^{n=6} n_i = N$$

freq. esperadas $(f_i^{\text{esp}} = Np_i = \epsilon_i) \quad \rightarrow \{\epsilon_1, \ldots, \epsilon_6\}$

$$\sum_{i=1}^{n=6} \epsilon_i = N \underbrace{\sum_{i=1}^{n=6} p_i}_{1} = N \qquad p_i = \frac{1}{6}$$

Amostra de N lançamentos

i	n_i	ϵ_i	n_i^2
1	16	20	254
2	19	20	361
3	27	20	729
4	17	20	289
5	23	20	589
6	18	20	324
			2488



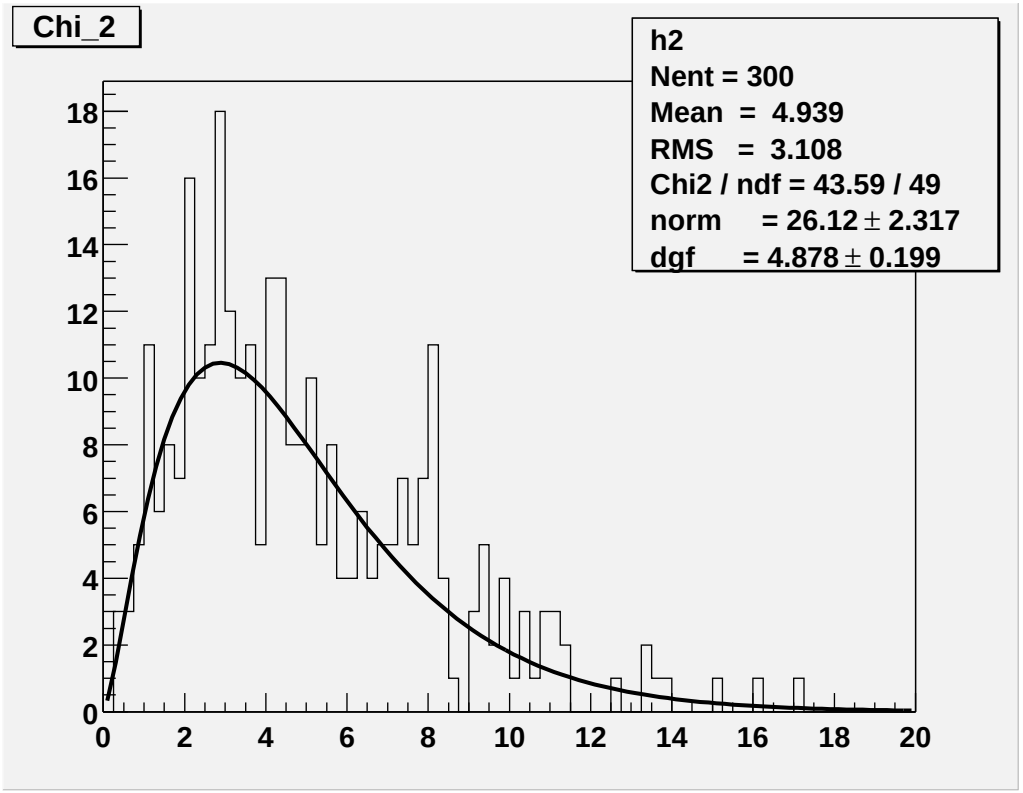
- As diferenças são significativas?
- Como caracterizar a discrepância?

Medidas de discrepância

- $(n_i - \epsilon_i) \neq 0$ (desvio)
- $\sum_{i=1}^{n}(n_i - \epsilon_i) = \underbrace{\sum_{i=1}^{n} n_i}_{N} - \underbrace{\sum_{i=1}^{n} \epsilon_i}_{N} = 0$
- $\sum_{i=1}^{n}(n_i - \epsilon_i)^2$ (não suficiente)
- $\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \epsilon_i)^2}{\epsilon_i} = \sum_{i=1}^{n} \frac{n_i^2}{\epsilon_i} - N = 4.4$

χ^2 depende de $(n-1)$ termos independentes
$\nu = n - 1$ (número de graus de liberdade)

Teste de χ^2



$$\begin{cases} \nu = 5 \\ \epsilon_i = N/6 \end{cases} \Longrightarrow \chi^2 = \frac{1}{N/6}\left(\sum_{i=1}^{n} n_i^2\right) - N = 4.4$$

Experimentos

$$\text{dados} = \{x_1, x_2, \ldots, x_N\} \quad \Longrightarrow \quad \mu_{\text{obs}} \text{ (resultado)}$$

Hipótese de nulidade

resultado esperado na ausência de qualquer outra causa que não seja o acaso.

$$\boxed{H_\circ : \mu = \mu_\circ}$$

$$\text{(discrepância)} \quad \mu_{\text{obs}} \neq \mu_\circ \quad \text{(real ou casual?)}$$

Fisher: o objetivo de todo experimento é testar evidências contra a hipótese de nulidade

Testes estatísticos

- pressupõem a escolha de uma distribuição de probabilidade (PDF), $\rho(\mu|\mu_\circ)$, para a distribuição dos dados ou para os valores de um parâmetro.
- testes de significância: consistência das observações (dados) com a hipótese de nulidade (não se considera hipóteses alternativas – Fisher)
- testes de hipótese: critérios para a rejeição ou aceitação da hipótese de nulidade (envolve hipóteses alternativas – Neyman-Pearson)

Fisher

- medida da evidência de que um valor observado de um parâmetro contrarie a hipótese de nulidade
- probabilidade de que ocorra um resultado maior do que o valor observado

$$\begin{cases} p = P(\mu \geq \mu_{\text{obs}}) = \displaystyle\int_{\mu_{\text{obs}}}^{\infty} \rho(\mu|\mu_\circ)\,\mathrm{d}\mu \quad (p-\text{value}) \\ \\ p \to 0 \quad (\text{valor observado não compatível com hipótese de nulidade}) \end{cases}$$

jargão estatístico

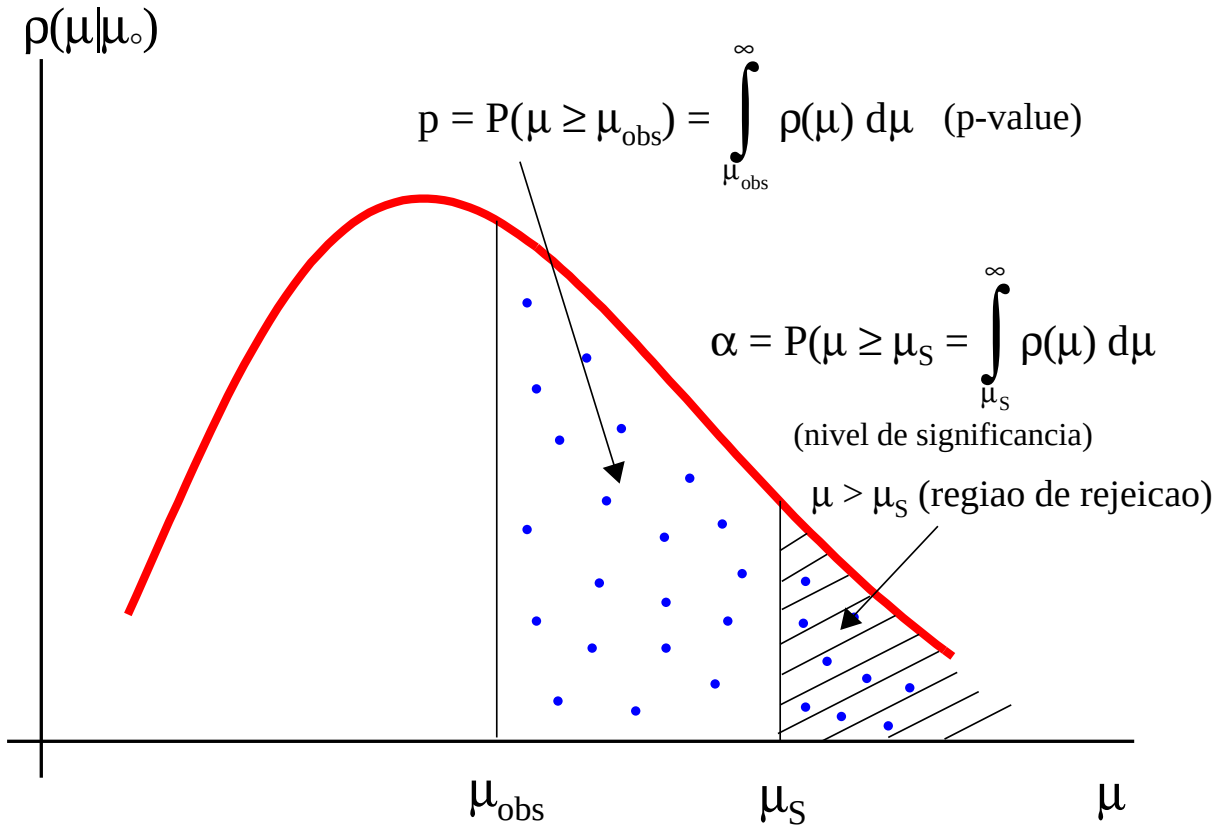
$0{,}01 < p < 0{,}05$ – diz-se que a evidência contra a hipótese de nulidade não é significativa ao nível de 1%, mas é significativa ao nível de 5%

Jerzy Neyman (1894-1981)

- prefixa arbitrariamente um **nível de significância** (α) e, consequentemente, um valor limite para qualquer valor observado
- $\alpha = P(\mu \geq \mu_S) = \int_{\mu_S}^{\infty} \rho(\mu|\mu_\circ)\,\mathrm{d}\mu$
- $\mu \geq \mu_S$ (região de rejeição)

- $\mu_{\mathrm{obs}} > \mu_S \iff p < \alpha$
- hipótese de nulidade é rejeitada ao nível de $100 \times \alpha\,\%$
- discrepância estatisticamente significativa ao nível de $100 \times \alpha\,\%$

p-value e nível de significância



$\alpha \to 1 \quad (\text{grande}) \iff$ grande região de rejeição

$\Downarrow$

probabilidade alta de rejeição

seja $H_\circ$ falsa ou verdadeira

rejeitar $H_\circ$ verdadeira $\implies$ falsa descoberta (erro tipo I)

- a rejeição de uma hipótese de nulidade não significa que $H_\circ$ seja realmente falsa, significa apenas que a hipótese não é compatível com os dados.
- outras hipóteses alternativas podem ser mais consistentes com os dados
- um único experimento não pode desacreditar uma "boa" hipótese (Fisher)

$\alpha \to 0$ (pequeno) $\iff$ grande região de aceitação

$\Downarrow$

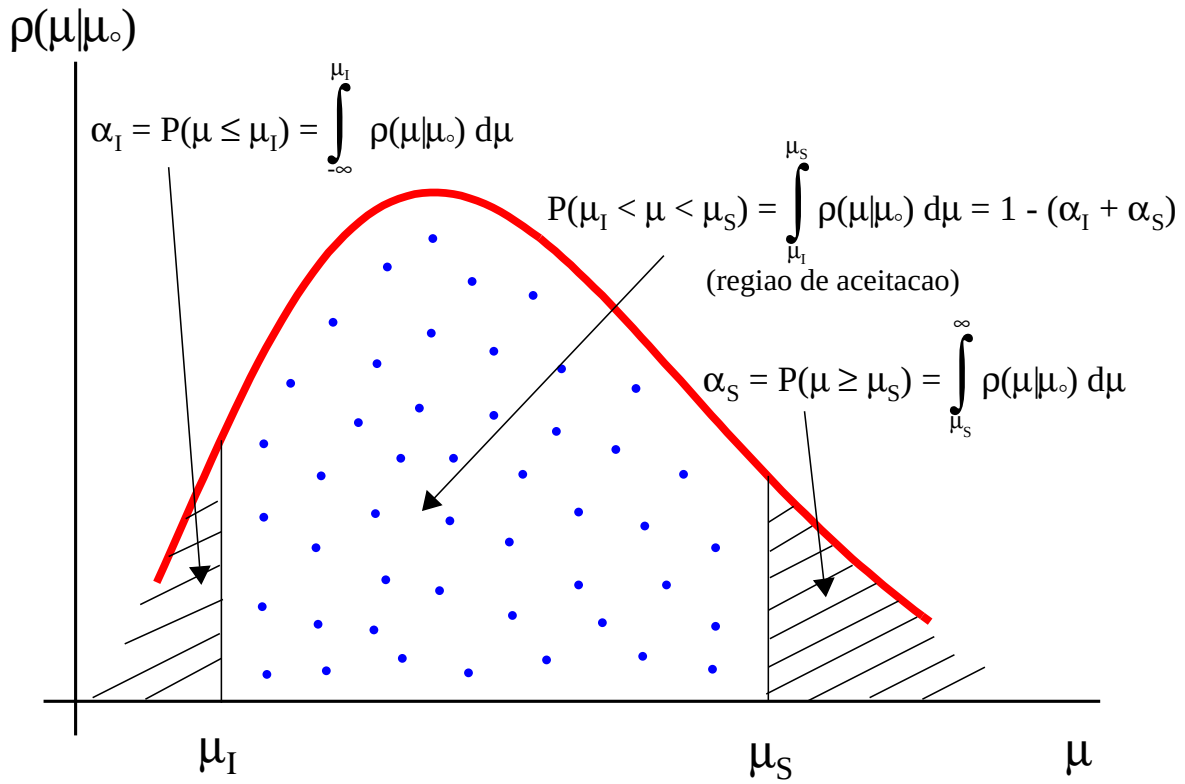
probabilidade alta de aceitação

seja $H_\circ$ falsa ou verdadeira

aceitar $H_\circ$ falsa $\implies$ refutar efeito real (erro tipo II)

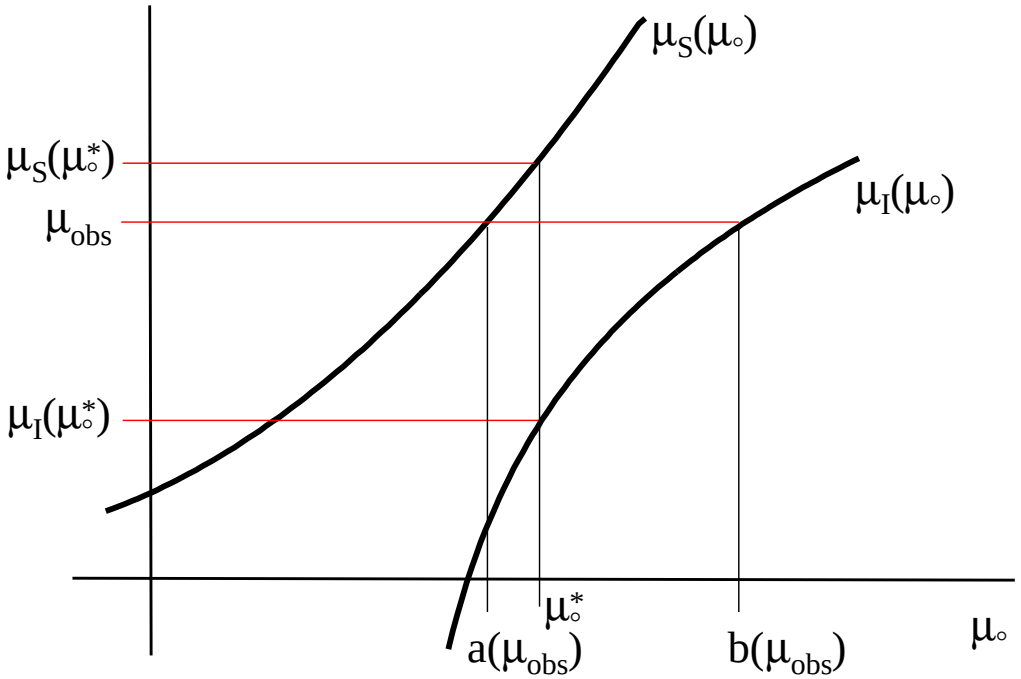
- a aceitação de uma hipótese de nulidade significa apenas que $H_\circ$ e os dados são estatisticamente consistentes
- outras hipóteses alternativas podem ser igualmente consistentes
- não é um único resultado que estabelece uma "prova" estatística

Nível de confiança



- $(\mu_I, \mu_S) \rightarrow$ região de aceitação da hipótese de nulidade ao **nível de confiança** (CL) de $[1 - (\alpha_I + \alpha_S)] \times 100$ %

Intervalo de confiança



- Um dado $\mu^*_\circ$ e um valor observado μ_{obs} só serão compatíveis se pertencerem, respectivamente, aos intervalos $[(a, b)$ e (μ_I, μ_S)

Intervalo de confiança

- para um dado valor esperado $\mu_\circ^*$ o valor observado μ_{obs} só estará em uma região de aceitação $\Big(\mu_I(\mu_\circ^*), \mu_S(\mu_\circ^*)\Big)$ definida por $\mu_\circ^*$ e pelo nível de confiança (CL) pré-fixado, se o intervalo $\Big(a(\mu_{\text{obs}}), b(\mu_{\text{obs}})\Big)$ determinado por μ_{obs} e pelo CL pré-fixado contiver o valor esperado $\mu_\circ^*$
- os intervalos (μ_I, μ_S) e (a, b) são probabilisticamente equivalentes, ou seja,
  $\mu_I(\mu_\circ^*) < \mu_{\text{obs}} < \mu_S(\mu_\circ^*) \iff a(\mu_{\text{obs}}) < \mu_\circ^* < b(\mu_{\text{obs}})$
- $P\big[\mu_I(\mu_\circ^*) < \mu_{\text{obs}} < \mu_S(\mu_\circ^*)\big] = P\big[a(\mu_{\text{obs}}) < \mu_\circ^* < b(\mu_{\text{obs}})\big]$
- (a, b) é um **intervalo de confiança** para $\mu_\circ$

Interpretações estatísticas

$$P(a < \mu_\circ < b)$$

- fração de vezes que o intervalo (aleatório) (a, b) conterá o valor esperado do parâmetro $\mu_\circ$ (fixo – não-aleatório) (frequentista – Pearson-Neyman)
- probabilidade de que o valor esperado do parâmetro $\mu_\circ$ (aleatório) ocorra no intervalo (a, b) (subjetiva – bayesiana)