APRENDIZADO DE MÁQUINAS PARA FÍSICA

Aula IV

Testes de Hipóteses e

Goodness-of-fit

TESTE DE HIPÓTESES

- Queremos decidir a concordância entre os dados experimentais e um (ou mais) modelo(s)
  - H_0 **hipótese nula** é testada, e a que se espera verdadeira
  - H_1 **modelo alternativo** que complementa H_0
- Se considerarmos H_0 como o *background* ou fundo e H_1 como o sinal, temos a matriz confusão:

realidade

decisão	Background (H_0)	Sinal (H_1)
Background(H_0)	Verdadeiro -	Falso -
Sinal(H_1)	Falso +	Verdadeiro +

erro tipo 2

erro tipo 1

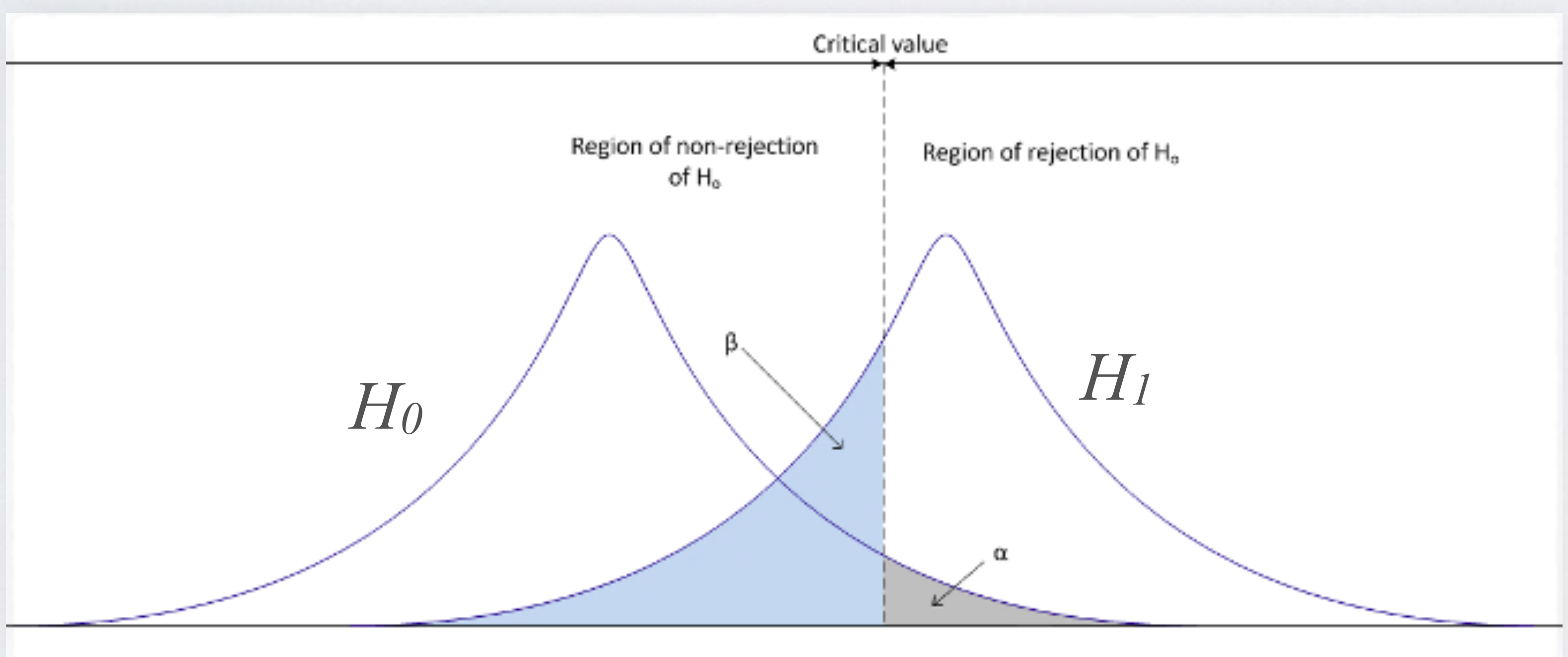
Para quantificar o nível de concordância dos dados com nossas hipóteses definimos um **teste estatístico** $t(\{x_i\})$ como uma **função escalar da amostra** $\{x_i\}$.
A utilidade do teste depende do seu poder de discriminação entre as hipóteses.

QUANTIFICANDO O TESTE

- Temos uma amostra de n medidas $\{x_i\}$. O teste estatístico é uma variável aleatória com PDFs $g(t|H_0)$ e $g(t|H_1)$.
  - ✳ A probabilidade de rejeitar H_0 sendo verdadeira $\alpha = \int_{t_c}^{\infty} g(t\,|\,H_0)dt$
  - ✳ A probabilidade de aceitar H_0 (e rejeitar H_1) sendo falsa $\beta = \int_{-\infty}^{t_c} g(t\,|\,H_1)dt$

erro tipo 1: “descoberta” falsa

erro tipo 2: “perder” a descoberta



α é a significancia
$1-\beta$ é o poder

Para testar a hipótese nula escolhemos um valor para α e a partir dele determinamos o valor crítico t_c. Logo avaliamos nos dados $t_{obs}=t(\{x_i\})$, e se $t_{obs}<t_c$ aceitamos H_0, caso contrário rejeitamos ela.

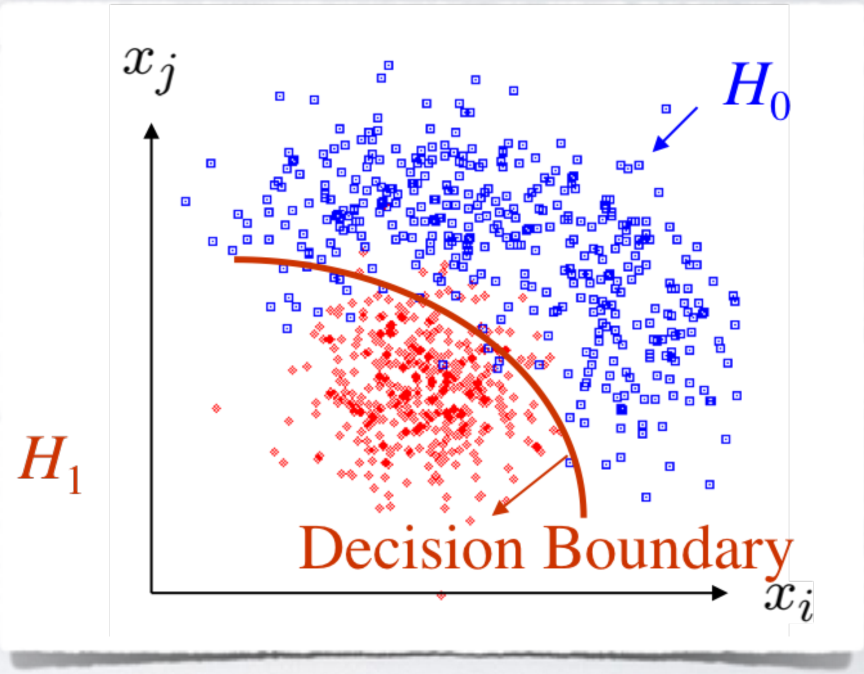
NEYMAN-PEARSON

- Existem muitas (infinitas!) possibilidades de estatística de teste ***t***.
- Likelihood ratio : a razão entre os valores de máxima verossimilhança para $\boldsymbol{H_0}$ e $\boldsymbol{H_1}$

$$t(\{x_i\}) = \frac{\mathscr{L}(H_1 | \{x_i\})}{\mathscr{L}(H_0 | \{x_i\})} \text{ ou equivalente } t(\vec{x}) = \frac{f(\vec{x} | H_1)}{f(\vec{x} | H_0)}$$

- O lemma de Neyman-Pearson diz que para um nível de significância α fixo, o LR é o teste mais poderoso (menor β)

Após escolhermos um valor de α e um teste estatístico, o valor crítico $t_c = t(\vec{x})$ define uma "fronteira" de decisão (hyper-superficie no espaço $\mathbb{R}^n$) que separa a região de aceitação e a de rejeição.



PROCEDIMENTO I

Se conhecemos $f(\vec{x}\,|H_0)$ and $f(\vec{x}\,|H_1)$ analiticamente os passos a seguir são:

1. determinar a PDF da estatística de teste no caso da hipótese nula $g(t\,|H_0)$
2. Definir um valor para α levando em consideração as probabilidades de error tipo 1 e 2, e obter a região crítica e fronteira.
3. determinar o valor de $t_{obs}=t(\{x_i\})$ para os dados
4. analisar se t_{obs} se encontra na região crítica (rejeitar H_0) ou na região de aceitação (não tem suficiente evidência para rejeitar H_0)

DEFINIÇÕES

Property of discriminator:

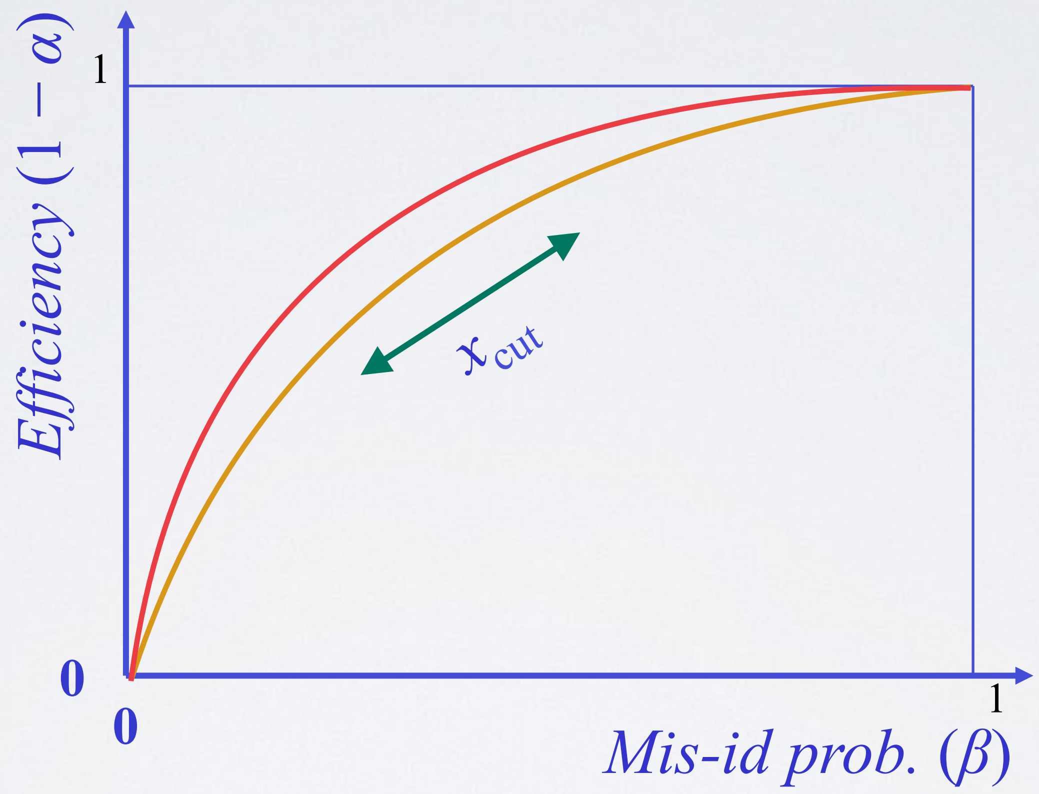
- Selection efficiency: probability to correctly identify signal events ε**=1-β**
- Misidentification probability: probability to misidentify as a background event **β**
- Purity: fraction of signal in a positively identified sample = $\mathbf{N_{V+} / N_{TOT+}}$
  - Depends on the signal and background composition! It is not a property of the discriminator only
- Fake rate: fraction of background in a positively identified sample, = 1 − Purity = $\mathbf{N_{F+} / N_{TOT+}}$

realidade

decisão	Background (H_0)	Sinal (H_1)
Background(H_0)	**background rejection** $\epsilon=1-\alpha$	β (misID)
Sinal(H_1)	α N_{V+}	**signal efficiency** $\varepsilon=1-\beta$ N_{T+}

$$N_{TOT+} = N_{T+} + N_{V+}$$

ANALISANDO A ESCOLHA DE t E O VALOR DO CORTE



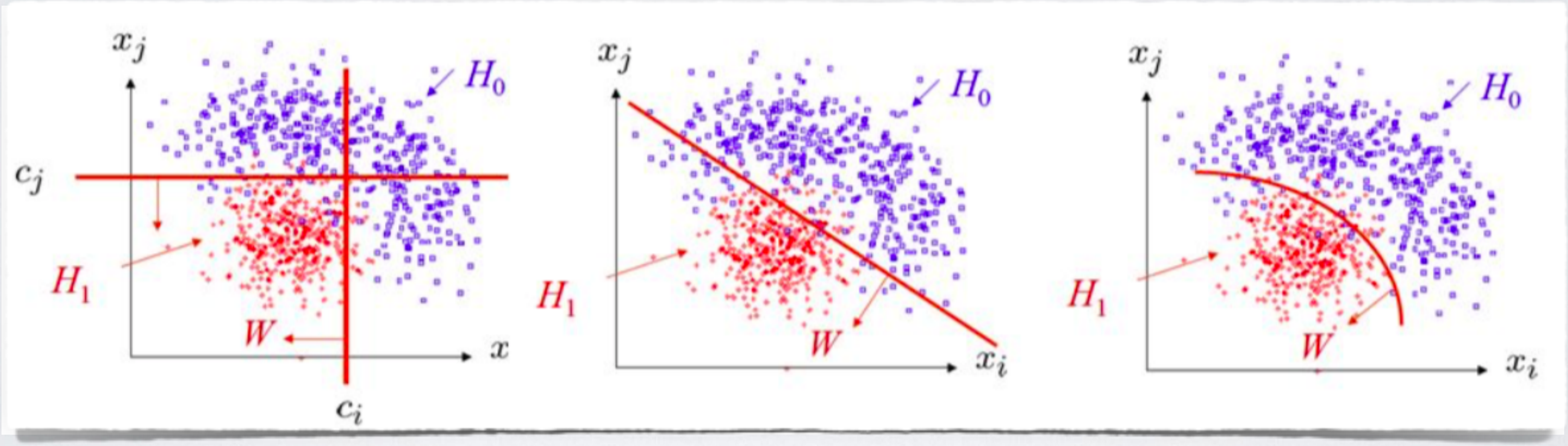
curvas ROC

Eficiência na identificação do background e taxa de falso negativo variam com a escolha do valor de "corte" (fronteira) no t. Um teste é melhor que outro para um mesmo valor de α se o valor de β é menor.

CONSTRUINDO t

Se não conhecemos as PDFs analiticamente, temos 2 opções:

- utilizar o método de Monte Carlo para gerar amostras e histogramas (em multiples dimensões) para H_0 e H_1. (problema se temos muitas variáveis xi)
- ou utilizar análise multi-variada para achar a fronteira (Fisher discriminant, BDT ou NN)

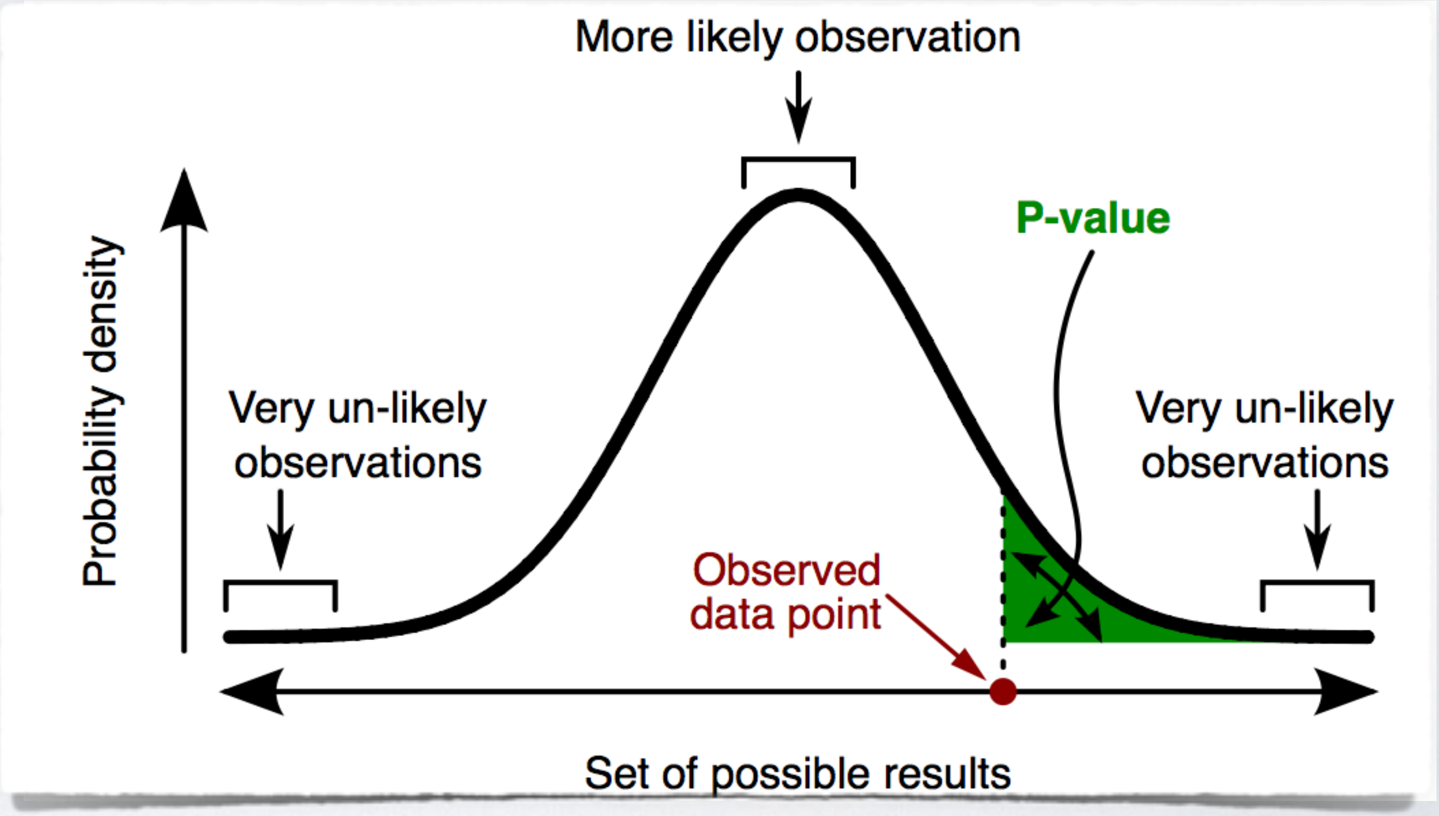


GOODNESS-OF-FIT

Se queremos quantificar a concordância de um teste com uma hipótese H_0 sem ter um modelo alternativo podemos avaliar o *goodness-of-fit*, quantificado pelo ***p*-value**: a probabilidade de se obter dados **tão ou mais** afastados da hipótese nula quanto os observados

$$p = \int_{t_{obs}}^{\infty} g(t \mid H_0)\,dt$$

- t_{obs} depende dos dados
- quanto menor o valor de p, maior a evidência **contra** H_0 (menor a chance de ser apenas flutuação)
- também chamado ***nível de significância observado***



TESTE DE HIPÓTESE VS GOF

- If one had defined a critical region, the significance level α of the hypothesis test would correspond to the *P-value*
- In hypothesis test α is a constant chosen a priori, before looking into data, while in GOF the observed *P-value* is a random variable
- In hypothesis test the critical value (boundary) t_c is a function of the chosen α, while in GOF it's the value of the test statistic evaluated on data t_{obs}
- Although *P-value* are not confidence levels, sometimes one uses the notations $CL_b = 1 - P_0$ and $CL_{s+b} = P_1$ in analyses, where P_0 and P_1 are *P-values* for $H_0(background)$ and $H_1(signal + background)$ hypotheses

SIGNIFICÂNCIA EM Z

- Frequentemente em vez de usar ***p-value*** alguns estudos utilizam o valor ***z:*** o número de desvios padrões que o ponto teria se uma distribuição normal padrão ($\mu = 0, \ \sigma = 1$) tiver a área equivalente na calda da direita

$$p = \sqrt{2\pi} \int_z^{\infty} e^{x^2/2} = \frac{1}{2}\left[1 - Erf\left(\frac{z}{\sqrt{2}}\right)\right]$$

p vem dos dados e a pdf , então z é achado invertendo a relação (ou tabelado)

- Em geral (em FAE) dizemos que temos uma observação quando $z \geq 3$ e uma evidência ou descobrimento se $z \geq 5$

EXEMPLO DIST. POISSON

- Probabilidade de uma observação de $n = n_s + n_b$ eventos com média esperada $\nu = \nu_s + \nu_b$

$$f(n \,|\, \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} \mathrm{e}^{-(\nu_s + \nu_b)}$$

- Significância observada seria a probabilidade de se observar $n \geq n_{obs}$ considerando a hipótese nula, quer dizer que seja só uma flutuação do background, sem sinal no modelo ($\nu_s = 0$)

$$P(n \geq n_{obs} \,|\, H_0 : \nu_s = 0) = \sum_{k=n_{obs}}^{\infty} P(k \,|\, H_0) = \sum_{k=n_{obs}}^{\infty} \frac{\nu_b^k}{k!} \mathrm{e}^{-\nu_b} = 1 - \sum_{k=0}^{n_{obs}-1} \frac{\nu_b^k}{k!} \mathrm{e}^{-\nu_b}$$

TESTE DO χ^2 DE PEARSON

- Se temos um histograma com N classes e n_i eventos por classe (bin) e no modelo esperamos ter ν_i eventos, a estatística de teste χ^2 de Pearson quantifica a concordância dos dados ao modelo

$$\chi^2 = \sum_i^N \frac{(n_i - \nu_i)^2}{\nu_i}$$

- Se os dados em cada bin seguem uma distribuição de Poisson com média (e variância)= ν_i e **n_i>5**, este teste segue a distribuição χ^2 com N variáveis e podemos avaliar o ***p-value***

$$p = P(z \geq \chi^2) = \int_{\chi^2}^{\infty} f(x|N)dx = 1 - \int_0^{\chi^2} f(x|N)dx$$

lembrando $$f(x|N) = \frac{1}{2^{N/2}\Gamma(N/2)} x^{(N/2-1)} \exp(-x/2)$$

⇒ dado um valor de χ^2 observado nos dados para um determinado N, podemos achar (tabelado, p.ex.) o ***p-value***.