Universidade de São Paulo
Instituto de Matemática e Estatística
Aluna: Nara Avila Moraes
Nro. USP: 5716734

2o. semestre, 2025
Prova - 04/11

MAE5911/IME: Fundamentos de Estatística e Machine Learning.
Prof.: Alexandre Galvão Patriota

Questão 01: Defina formalmente o valor-p (**aproximado**) para testar uma hipótese geral e discuta os problemas de interpretação usuais ao utilizar a sua versão condicional à hipótese nula. Apresente um exemplo numérico e simulações de Monte Carlo, usando um modelo de regressão Poisson, para ilustrar a sua aplicação.

Questão 02: Apresente um Teorema da Aproximação Universal que generalize o Teorema de Funahashi, discuta as diferenças entre os resultados por meio de exemplos de redes neurais.

Questão 03: Proponha uma função de estimação robusta para estimar os parâmetros de uma regressão Binomial, $Y|X = x \sim \text{Bin}(m, \mu_\theta(x))$, em que $\mu_\theta(\cdot)$ é uma função com imagem em $[0, 1]$ e θ é o vetor de parâmetros. Apresente um exemplo numérico e uma simulação de Monte Carlo, como feito em sala, para ilustrar os resultados. Perturbe a distribuição dos dados e mostre que a sua proposta é de fato robusta contra essa perturbação.

Resposta para o Questão 01

O conceito do p-valor, tal como proposto originalmente por *Fisher* (1925), é uma medida **condicional sob** H_0, definida por

$$\text{valor-p}(H_0, z_n) = \sup_{\theta \in \Theta_0} P_\theta^{(n)}\big(T_{H_0}(Z_n) \geq T_{H_0}(z_n)\big),$$

isto é, a probabilidade de se observar uma estatística tão extrema quanto a obtida, assumindo H_0 verdadeira. O valor-p foi concebido por Fisher como um *instrumento de contraste* entre o dado e hipótese, mas não como uma medida de verdade. Seu objetivo era expressar o grau de incompatibilidade empírica entre o fenômeno observado e as consequências lógicas de H_0.

Posteriormente, a escola *Neyman–Pearson* reinterpretou esse conceito dentro de uma estrutura de decisão. Nessa transição, o valor-p perdeu seu caráter exploratório e passou a ser tratado como um critério binário, externo e supostamente não condicionado a H_0, capaz de expressar a decisão: "rejeita" ou "não rejeita" H_0. Esta leitura gerou um equívoco lógico: o valor-p, que é definido dentro do universo em que H_0 é verdadeira, passou a ser interpretado como se pudesse mensurar a probabilidade de H_0 ser verdadeira.

O **valor-p assintótico**, que é o valor usualmente utilizado, na ausência da distribuição exata de T_{H_0}, é definido por:

$$p^{(a)}(H_0, z_n) = \sup_{\theta \in \Theta_0} P_\theta(W_\theta \geq T_{H_0}(z_n)),$$

em que W_θ representa a distribuição-limite de T_{H_0} quando $n \to \infty$.

Confirmando essa natureza epistêmica, o valor-p assintótico nem sempre converge para o valor-p exato, isto porque, estando definido condicionalmente a H_0, pode não ser sensível a falhas estruturais do modelo, como a ausência de independência entre as observações, a presença de hipóteses condicionais aninhadas, ou a violação das condições de regularidade que garantem a validade assintótica — como homocedasticidade.

Por isso, o valor-p deve ser entendido apenas como originalmente proposto — uma **medida de discrepância entre o real e o teórico** — que opera em um único sentido: como evidência contra a hipótese nula, mas nunca para confirmá-la.

Em termos das σ-álgebras envolvidas, a construção inferencial pode ser representada como uma sequência de aplicações mensuráveis, onde cada espaço de probabilidade possui a sua própria σ-álgebra, refletindo níveis distintos de abstração:

$$(\Omega_n, \mathcal{F}_n, \mathbb{P}_n) \xrightarrow{Z_n} (\mathcal{Z}_n, \mathcal{B}_{Z_n}, \mathbb{P}_{Z_n}) \xrightarrow{T} (\mathcal{T}_n, \mathcal{B}_{T_n}, \mathbb{P}_{T_n}).$$

- $\mathcal{F}_n$ — descreve o universo empírico de incertezas, isto é, os eventos do mundo observável;

- $\mathcal{B}_{Z_n}$ — corresponde ao espaço amostral modelado, onde as variáveis aleatórias Z_n são definidas sob $\mathbb{P}_{Z_n}$;
- $\mathcal{B}_{T_n}$ — é a σ-álgebra induzida pela estatística T, onde vivem as distribuições teóricas dos testes e, em particular, o valor-p.

O valor-p pertence ao universo lógico em que a hipótese nula H_0 é assumida como verdadeira. Ele não pertence à σ-álgebra empírica $\mathcal{F}_n$, e portanto não pode ser interpretado como uma probabilidade sobre o mundo real ou sobre a veracidade de H_0. Tratá-lo dessa forma constitui um erro de violação da σ-álgebra — um deslize epistemológico em que se tenta extrair informação empírica de uma construção em sua definição epistêmica condicional e autoreferente.

Implementações

O exemplo numérico apresentado na Listing 2 tem o objetivo de ilustrar o comportamento do valor-p sob diferentes hipóteses nulas, mantendo fixos os dados observados. O experimento baseia-se em um modelo de regressão de Poisson definido por

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{com} \quad \log(\lambda_i) = \theta_0 + \theta x_i,$$

onde $\theta_0 = 1$ e o valor verdadeiro do parâmetro é $\theta = 0{,}20$. Uma única amostra de tamanho $n = 200$ foi gerada segundo esse modelo, representando o conjunto de dados "observados" em um mundo empírico fixo.

A partir dessa amostra, foram testadas três hipóteses nulas distintas sobre o mesmo conjunto de dados:

$$H_0^{(A)} = 0{,}20, \qquad H_0^{(B)} = 0{,}10, \qquad H_0^{(C)} = 0{,}60.$$

Cada hipótese foi avaliada por meio da estatística de Wald,

$$Z = \frac{\hat{\theta} - \theta_{H_0}}{\text{se}(\hat{\theta})},$$

e o valor-p correspondente foi calculado de forma:

$$p = P(Z \leq z_{\text{obs}}),$$

correspondendo à hipótese alternativa $H_1 : \theta < \theta_{H_0}$.

Os resultados obtidos encontram-se na Tabela 1. Observa-se que, quando a hipótese nula coincide com o valor real do parâmetro (θ), o valor-p é elevado e a hipótese não é rejeitada. À medida que o valor hipotético se afasta de θ, o valor-p decresce rapidamente, levando à rejeição de H_0 nos casos de maior discrepância ($H_0^{(C)}$). Essa simulação evidencia o papel do valor-p como medida de compatibilidade entre modelo e dado: ele quantifica o grau de discrepância entre a hipótese estatística e a estrutura empírica observada.

Tabela 1: Resultados do cálculo do valor-p para diferentes hipóteses nulas sob o modelo de regressão de Poisson (teste unicaudal à esquerda).

Cenário	**Hipótese nula**	**Valor-p**	**Decisão**	$\hat{\theta}$	θ
Cenário A ($H_0^{(A)}$ exato)	$\theta_{H_0} = 0{,}20$	0,787386	Não rejeita H_0	0,254359	0,20
Cenário B ($H_0^{(B)}$ com pequeno desvio)	$\theta_{H_0} = 0{,}28$	0,353412	Não rejeita H_0	0,254359	0,20
Cenário C ($H_0^{(C)}$ com grande desvio)	$\theta_{H_0} = 0{,}60$	0,000000	Rejeita H_0	0,254359	0,20

O experimento de Monte Carlo tem como objetivo avaliar, de forma empírica, o comportamento assintótico do valor-p quando múltiplas amostras são geradas a partir de um mesmo modelo de regressão de Poisson. Considera-se o modelo

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{com} \quad \log(\lambda_i) = \theta_0 + \theta x_i,$$

onde $\theta_0 = 1$ e o valor verdadeiro do parâmetro é $\theta = 0{,}20$. Em cada replicação, é gerada uma amostra de tamanho $n = 200$ com covariáveis $x_i \sim \text{Unif}(0, 2)$ e respostas Y_i conforme o modelo acima. O processo é repetido $M = 10{,}000$ vezes, mantendo θ fixo.

Para cada amostra, ajusta-se um modelo de regressão de Poisson e testam-se três hipóteses nulas distintas sobre o mesmo conjunto de dados:

$$H_0^{(A)} : \theta = 0{,}20, \qquad H_0^{(B)} : \theta = 0{,}10, \qquad H_0^{(C)} : \theta = 0{,}60.$$

Cada hipótese é avaliada pela estatística de Wald,

$$Z = \frac{\hat{\theta} - \theta_{H_0}}{\text{se}(\hat{\theta})},$$

e o valor-p correspondente é calculado de forma:

$$p = P(Z \geq z_{\text{obs}}).$$

A partir das M replicações, calcula-se a proporção de rejeição de cada hipótese, a média e a mediana dos valores-p, bem como a média dos estimadores $\hat{\theta}$. Os resultados são apresentados na Tabela 2.

Observa-se que, quando $H_0^{(A)}$ coincide com o valor verdadeiro do parâmetro, a proporção de rejeição aproxima-se do nível $\alpha = 0{,}05$, e os valores-p seguem aproximadamente uma distribuição uniforme. Quando a hipótese nula se afasta do valor real — como em $H_0^{(B)}$ e $H_0^{(C)}$ — a potência do teste aumenta, os valores-p concentram-se próximos de zero e a rejeição de H_0 torna-se praticamente certa em cenários de grande discrepância. Essa simulação evidencia o comportamento esperado do teste de Wald sob o modelo de Poisson e reforça o valor-p como medida de compatibilidade entre hipótese estatística e estrutura empírica observada.

Tabela 2: Resultados da simulação de Monte Carlo para o teste de Wald em regressão de Poisson. Cada linha apresenta os resultados obtidos para uma hipótese nula distinta, com $M = 10{,}000$ replicações.

Cenário	**Hipótese nula**	**% de rejeição**	**Média(p)**	**Mediana(p)**	$p_{0,05}$	$p_{0,95}$	$\bar{\hat{\theta}}$ / θ
Cenário A	$H_0^{(A)} = 0{,}20$	0,047700	0,497509	0,491766	0,052861	0,946414	0,200769 / 0,2
Cenário B	$H_0^{(B)} = 0{,}10$	0,441300	0,143984	0,066063	0,000982	0,543902	0,200769 / 0,2
Cenário C	$H_0^{(C)} = 0{,}60$	0,000000	0,999978	1,000000	0,999989	1,000000	0,200769 / 0,2

Nas simulações de Monte Carlo realizadas sob H_0, os p-valores mostraram-se aproximadamente uniformes, reproduzindo o comportamento esperado quando as condições assintóticas são satisfeitas. No entanto, ao introduzir pequenas violações, como superdispersão ou dependência entre observações, a distribuição dos p-valores tornou-se assimetricamente inclinada, evidenciando a sua sensibilidade à estrutura do modelo.

Esse resultado empírico confirma que a discrepância entre o valor-p exato e o assintótico não é meramente numérica, mas reflete o descompasso lógico entre o universo ideal de H_0 e o comportamento real dos dados.

```
# =====================================================================
# Regressão Poisson (GLM)
# =====================================================================

set.seed(111)

# Parâmetros fixos do "mundo real"
n      = 100
beta0  = 1.0
theta  = 0.20     # valor verdadeiro
alpha  = 0.05

# Covariável e única amostra observada
x       = runif(n, 0, 2)
```

```r
lambda  = exp(beta0 + theta * x)
y       = rpois(n, lambda)

# Ajuste único do GLM Poisson nos MESMOS dados
modelo  = glm(y ~ x, family = poisson)

# ---------------------------------------------------------------
# Função: testa H0: theta = h0 (Wald)
# ---------------------------------------------------------------
testa_glm_theta = function(mod, h0, alpha = 0.05) {
  theta_hat = coef(mod)[2]
  se_hat    = sqrt(vcov(mod)[2, 2])
  Z         = (theta_hat - h0) / se_hat

  # valor-p unicaudal (contra H1: theta > h0)
  p_valor   = pnorm(Z)

  decisao   = ifelse(p_valor < alpha, "Rejeita H0", "Não rejeita H0")

  data.frame(
    hipotese_nula = paste0("H0: theta = ", format(h0, nsmall = 2)),
    valor_p       = p_valor,
    decisao       = decisao,
    theta_hat     = theta_hat,
    theta         = theta
  )
}

# Hipóteses nulas a testar no MESMO conjunto (exato, pouco e muito desvio)
H0_A = 0.20   # H0 exato
H0_B = 0.28   # H0 com pouco desvio
H0_C = 0.60   # H0 com muito desvio

resA = testa_glm_theta(modelo, H0_A, alpha)
resB = testa_glm_theta(modelo, H0_B, alpha)
resC = testa_glm_theta(modelo, H0_C, alpha)

# Consolida com rbind (mantendo os rótulos de cenário como nomes de linha)
resumo_glm = rbind(
  "Cenário A (H0 exato)"        = resA,
  "Cenário B (H0 pouco desvio)" = resB,
  "Cenário C (H0 muito desvio)" = resC
)

# -------- Formatação (sem notação científica e 6 casas decimais) --------
options(scipen = 999)

resumo_fmt           = resumo_glm
resumo_fmt$valor_p   = format(round(resumo_fmt$valor_p,   6), nsmall = 6)
resumo_fmt$theta_hat = format(round(resumo_fmt$theta_hat, 6), nsmall = 6)
resumo_fmt$theta     = format(round(resumo_fmt$theta,     6), nsmall = 6)

# Tabela final (analítica)
noquote(resumo_fmt)
```

Listing 1: Exemplo numérico - Regressão Poisson

```r
set.seed(111)

# -------------------------
# Parâmetros do experimento
# -------------------------
n      = 200        # tamanho da amostra
beta0  = 1.0        # intercepto fixo
theta  = 0.20       # (verdadeiro) = coef. de x, usado para gerar os dados
alpha  = 0.05       # nível
M      = 10000      # repetições de Monte Carlo

# hipóteses nulas a testar (todas no mesmo conjunto de dados por replicação)
```

```r
H0_A = 0.20   # H0 exato
H0_B = 0.28   # H0 com pouco desvio
H0_C = 0.60   # H0 com grande desvio
H0s  = c(H0_A, H0_B, H0_C)

# ------------------------------------------
# Design fixo: x é gerado 1 vez e mantido fixo
# ------------------------------------------
x = runif(n, 0, 2)

# ------------------------------------------------------------
# Função: calcula p-valor unicaudal (esquerda) e decisão para b0
# ------------------------------------------------------------
p_unicaudal_wald = function(mod, b0, alpha) {
  theta_hat = coef(mod)[2]
  se_hat    = sqrt(vcov(mod)[2, 2])
  Z         = (theta_hat - b0) / se_hat
  # unicaudal à esquerda: H1: theta < b0
  p_valor   = pnorm(Z)
  decisao   = ifelse(p_valor < alpha, "Rejeita H0", "Não rejeita H0")
  list(p = p_valor, decisao = decisao, theta_hat = theta_hat)
}

# ------------------------------------------------------------
# Uma replicação: gera y ~ Poisson(exp(beta0 + theta * x)), ajusta GLM
# e testa H0_A, H0_B, H0_C na mesma amostra
# ------------------------------------------------------------
uma_replicacao = function() {
  lambda = exp(beta0 + theta * x)
  y      = rpois(n, lambda)
  mod    = glm(y ~ x, family = poisson)
  lapply(H0s, function(b0) p_unicaudal_wald(mod, b0, alpha))
}

# ------------------------------------------------------------
# Roda M replicações e consolida resultados por cenário
# ------------------------------------------------------------
# estruturas de acumulação
p_list_A = numeric(M); dec_A = character(M); thA = numeric(M)
p_list_B = numeric(M); dec_B = character(M); thB = numeric(M)
p_list_C = numeric(M); dec_C = character(M); thC = numeric(M)

for (m in 1:M) {
  out = uma_replicacao()
  # A
  p_list_A[m] = out[[1]]$p
  dec_A[m]    = out[[1]]$decisao
  thA[m]      = out[[1]]$theta_hat
  # B
  p_list_B[m] = out[[2]]$p
  dec_B[m]    = out[[2]]$decisao
  thB[m]      = out[[2]]$theta_hat
  # C
  p_list_C[m] = out[[3]]$p
  dec_C[m]    = out[[3]]$decisao
  thC[m]      = out[[3]]$theta_hat
}

# ------------------------------------------------------------
# Sumários: proporção de rejeição (potência/erro tipo I),
# estatísticas de p e média de theta_hat
# ------------------------------------------------------------
resumo_mc = data.frame(
  hipotese_nula = c(
    paste0("H0_A: theta = ", format(H0_A, nsmall = 2)),
    paste0("H0_B: theta = ", format(H0_B, nsmall = 2)),
    paste0("H0_C: theta = ", format(H0_C, nsmall = 2))
  ),
```

```r
  prop_rejeicao = c(
    mean(dec_A == "Rejeita H0"),
    mean(dec_B == "Rejeita H0"),
    mean(dec_C == "Rejeita H0")
  ),
  media_p  = c(mean(p_list_A), mean(p_list_B), mean(p_list_C)),
  mediana_p= c(median(p_list_A), median(p_list_B), median(p_list_C)),
  q05_p    = c(quantile(p_list_A, 0.05), quantile(p_list_B, 0.05), quantile(p_list_C, 0.05)),
  q95_p    = c(quantile(p_list_A, 0.95), quantile(p_list_B, 0.95), quantile(p_list_C, 0.95)),
  theta_hat_medio = c(mean(thA), mean(thB), mean(thC)),
  theta_real      = theta
)

# formatação
options(scipen = 999)
resumo_mc$prop_rejeicao   = format(round(resumo_mc$prop_rejeicao, 6), nsmall = 6)
resumo_mc$media_p         = format(round(resumo_mc$media_p,       6), nsmall = 6)
resumo_mc$mediana_p       = format(round(resumo_mc$mediana_p,     6), nsmall = 6)
resumo_mc$q05_p           = format(round(resumo_mc$q05_p,         6), nsmall = 6)
resumo_mc$q95_p           = format(round(resumo_mc$q95_p,         6), nsmall = 6)
resumo_mc$theta_hat_medio = format(round(resumo_mc$theta_hat_medio, 6), nsmall = 6)
resumo_mc$theta_real      = format(round(resumo_mc$theta_real,      6), nsmall = 6)

# Impressao da tabela
rownames(resumo_mc) = c(
  "Cenário A (H0_A exato)",
  "Cenário B (H0_B pouco desvio)",
  "Cenário C (H0_C muito desvio)"
)

noquote(resumo_mc)
```

Listing 2: Exemplo numérico - Regressão Poisson

Resposta para o Questão 02

Resposta para o Questão 03

Os códigos desde estudo estão disponibilizados em <http://academic-codex.github.io/MAE5911-Estatistica-e-Machine-Learning>.