

BIE5781

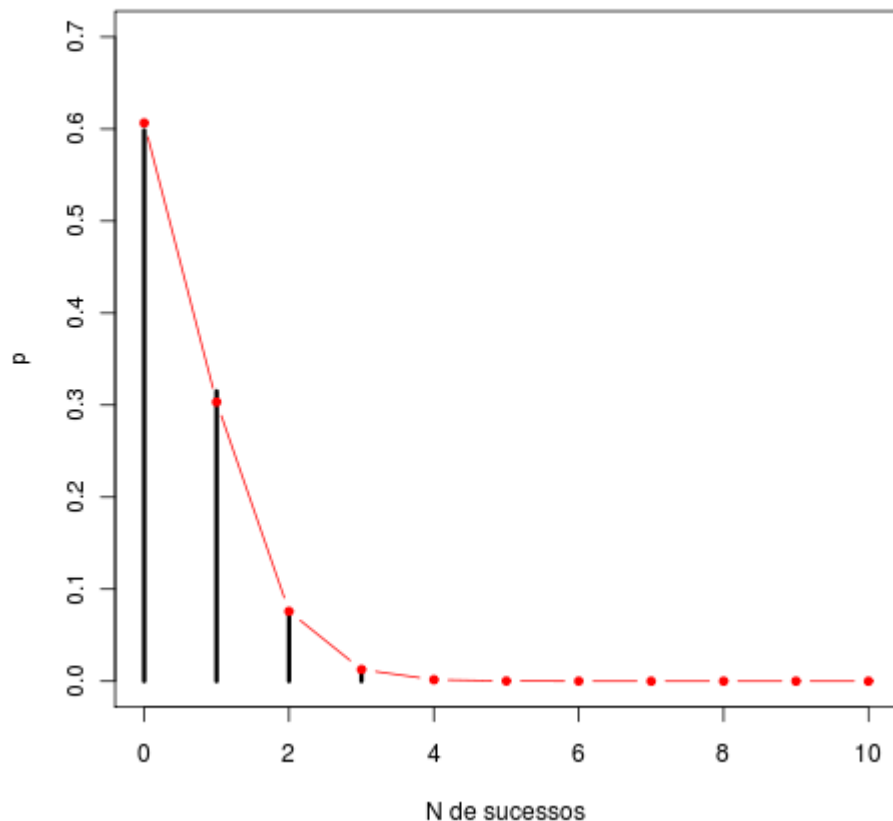
Aula 5

Seleção de Modelos

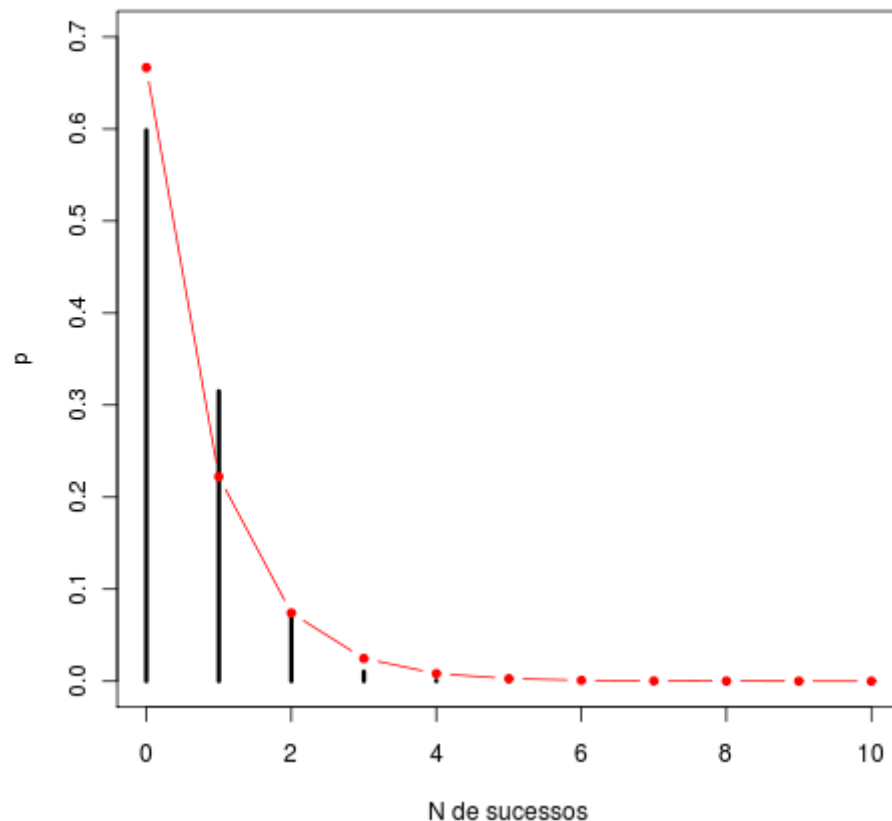


Como Medir a Diferença entre Distribuições?

X ~ Bin(N=10,p=0,05), Aproximação Poisson

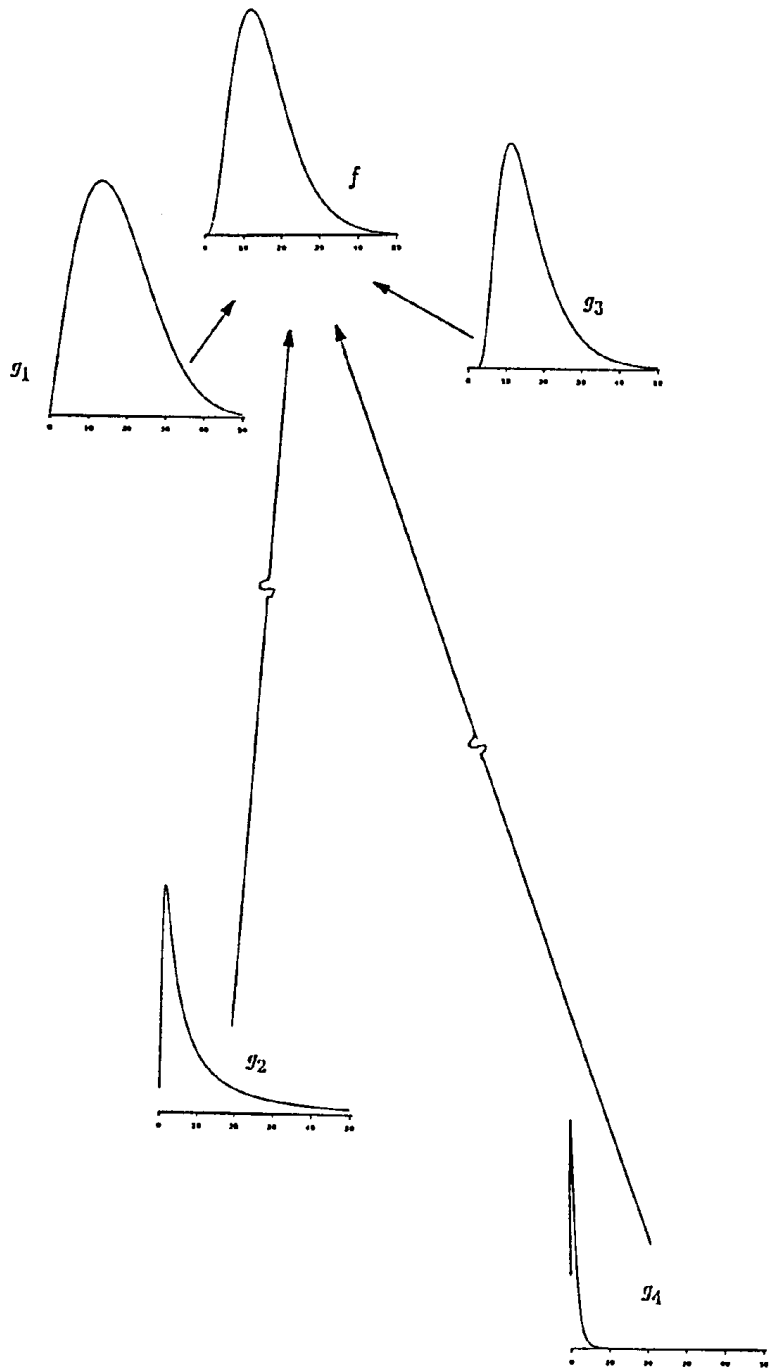


X ~ Bin(N=10,p=0,05), Aproximação Geometrica



$$I(f, g) = \sum f(x_i) \ln \frac{f(x)}{g(x_i|\theta)}$$

Distância de Kullback-Leibler



$$I(f, g) = \int f(x) \ln \frac{f(x)}{g(x|\theta)}$$



Ludwig Boltzmann
1844-1906



Solomon Kullback
1907-1994



Richard Leibler
1914-2003

Distância Relativa ao Modelo "Verdadeiro"

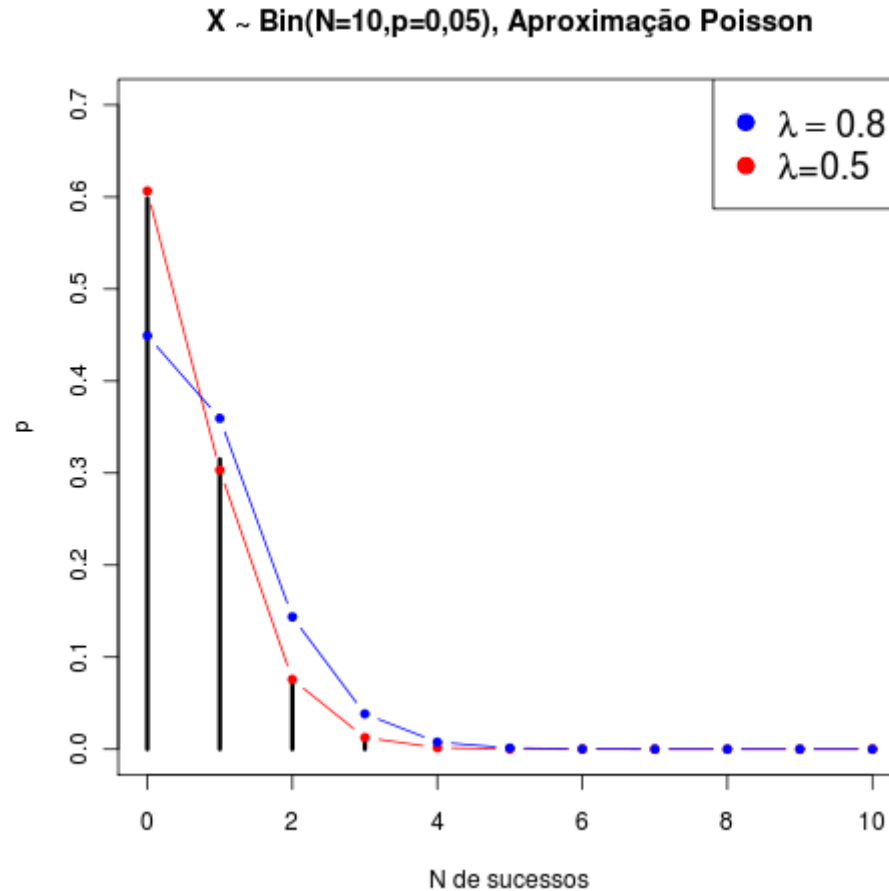
$$\begin{aligned} I(f, g) &= \sum f(x_i) \ln \frac{f(x)}{g(x_i|\theta)} \\ &= \sum f(x_i) \ln f(x) - \sum f(x_i) \ln g(x_i|\theta) \\ &= C - \sum f(x_i) \ln g(x_i|\theta) \\ &= C - E_x[\ln g(x_i|\theta)] \end{aligned}$$

DKL relativa: Exemplo

x	f(x)	g(x)	ln g(x)	ln g(x) f(x)
0	0.599	0.607	-0.500	0.299
1	0.315	0.303	-1.193	0.376
2	0.075	0.076	-2.579	0.193
3	0.010	0.013	-4.371	0.046
4	0.001	0.002	-6.451	0.006
5	0.000	0.000	-8.753	0.001
6	0.000	0.000	-11.238	0.000
7	0.000	0.000	-13.877	0.000
8	0.000	0.000	-16.650	0.000
9	0.000	0.000	-19.540	0.000
10	0.000	0.000	-22.536	0.000

$$\sum_{i=1}^{10} f(x_i) \ln(g(x_i|\theta)) = 0,921$$

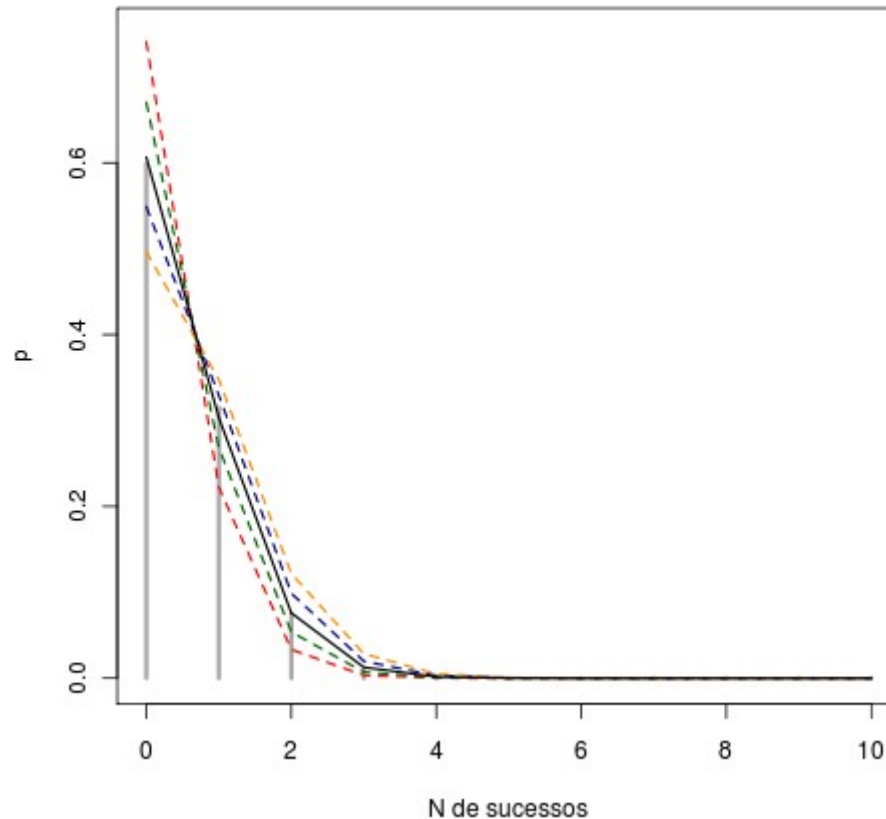
A Dura Realidade



$$\hat{I}(f, g) = C - \sum f(x_i) \ln g(x_i | \hat{\theta}(y))$$

Uma Dupla Esperança

$X \sim \text{Bin}(N=10, p=0,05)$, Aprox. Poisson

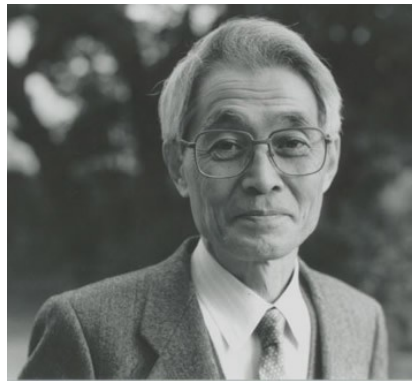


$$T = E_y E_x [\ln g(x_i | \hat{\theta}(y))] \\ \hat{T} = \ln L(\theta | y) - K$$

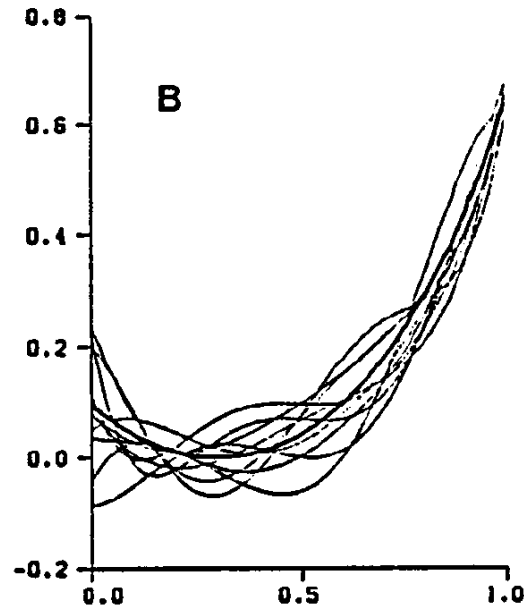
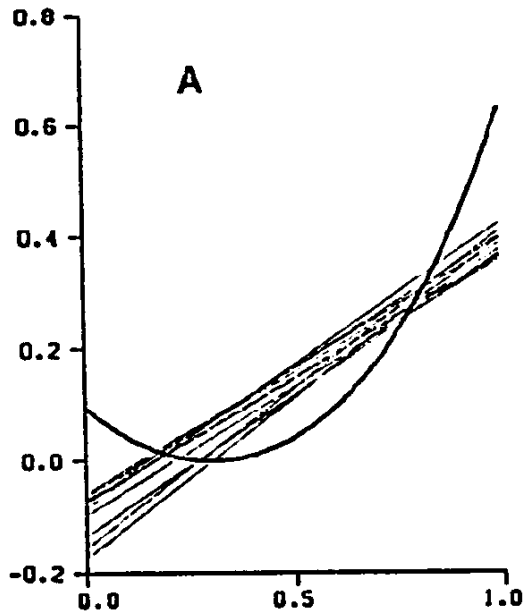
AIC

$$AIC = -2 \ln L(\theta|y) + 2K$$

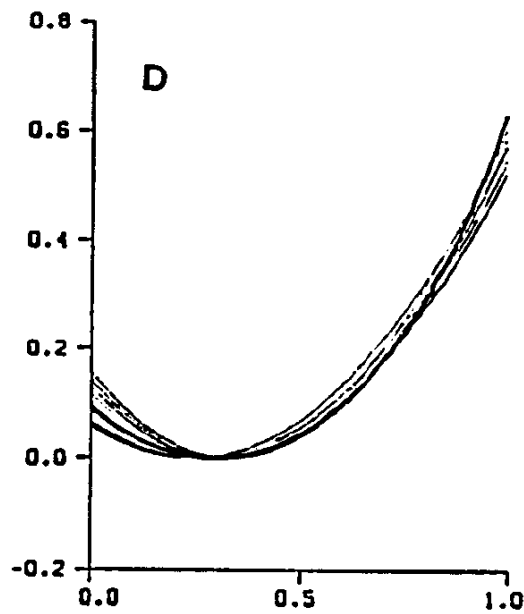
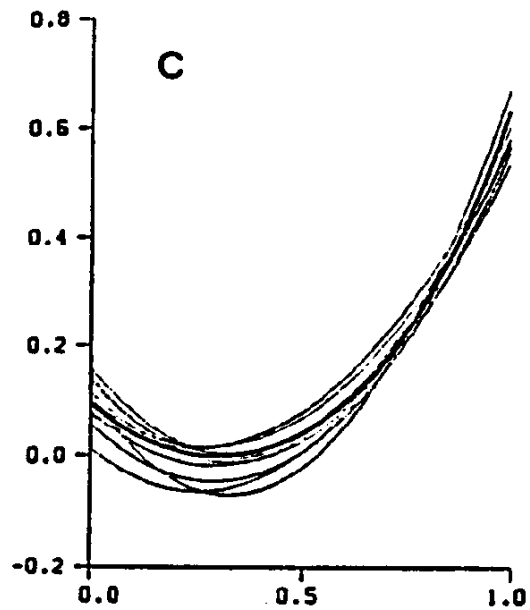
AIC = - 2 x Ln(verossimilhança) + 2 x n de parâmetros



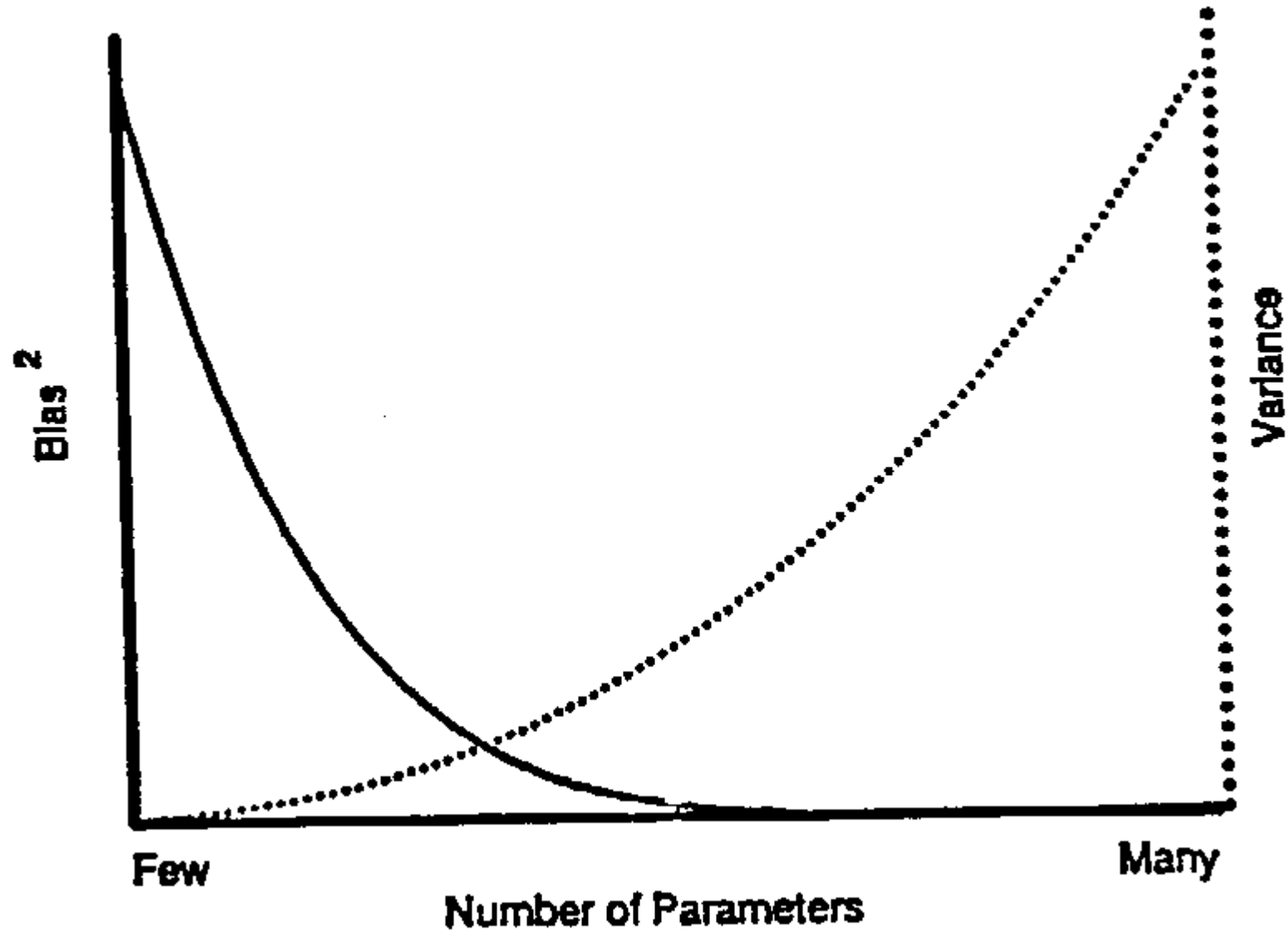
Hirotugu Akaike (1927)



Viés
X
Variância



O Balanço entre viés e variância



RESUMINDO

- Como não há modelo verdadeiro, comparamos nossos modelos em termos relativos.
- O AIC mostra que a verossimilhança pode ser usada como medida de discrepância relativa.
- O AIC não é apenas uma verossimilhança penalizada pelo número de parâmetros (embora esta seja uma propriedade conveniente dele).