# *Modelos estatísticos para previsão de partidas de futebol*

**Dani Gamerman**
**Instituto de Matemática, UFRJ**
**dani@im.ufrj.br**

**X Semana da Matemática e II Semana da Estatística da UFOP**

**Ouro Preto, MG – 03/11/2010**

## Algumas perguntas que queremos responder:

Resultados dos jogos futuros;

Quantos pontos serão necessários para se garantir o Cruzeiro na Libertadores;

Quantos pontos serão necessários para o Galo se livrar do rebaixamento;

Quantos pontos serão necessários para ganhar o título;

Quais as chances do Flamengo terminar na frente do Vasco.

Qual a colocação do Fluminense?

Muitos grupos de pesquisa tratando disso

Casas de apostas (virtuais) usam estatísticos

Tratamento científico deu origem a várias publicações

Grupos fazendo isso no Brasil:

Mat/UFMG – Bernardo Lima e co-autores
Est/UFSCar – Francisco Louzada e co-autores
Est/UFF – Leonardo Bastos e co-autores
Est/USP – Marcelo Arruda (chancedegol.com.br)
etc...

# Fatos estilizados:

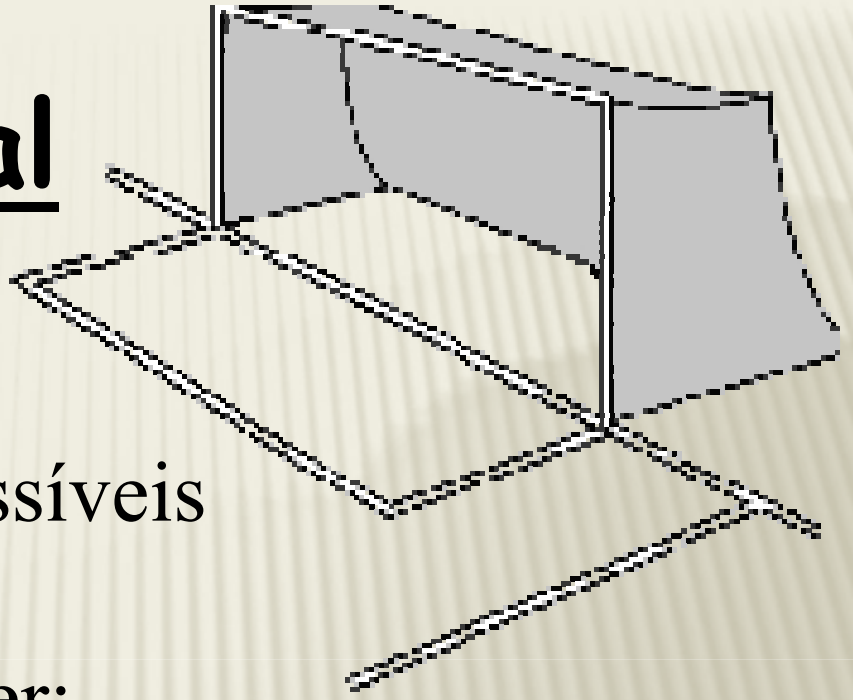
- futebol é um dos esportes mais incertos;
  um dos poucos onde o pior pode ganhar

- incerteza quantificada com probabilidades

- não se pode dizer nada com alta probabilidade;
  muito menos com rodadas de antecedência

- requer tratamento da dependência temporal entre rodadas do campeonato

# Espaço amostral



Conjunto de resultados possíveis

Para cada jogo, podemos ter:

- vitória, empate e derrota
- número de gols de cada time

Probabilistas trabalham com (vit,emp,der)
atualizadas segundo *técnica* de alisamento exponencial

Estatísticos trabalham com # de gols

# de gols é uma contagem

modelo natural é o Poisson

alguns usam técnica de alisamento
outros usam modelo sem tratar dependência temporal

Tratamento adequado deveria passar por

Formulação de um **modelo** estatístico

Forma científica de especificar (e testar) conjecturas

Incorporar todas as características do problema, especialmente a dependência temporal

# Como avaliar resultados?

Considere 3 preditores do clima: P1, P2 e P3.

P1 e P2 disseram que hoje ia fazer sol
P3 disse que hoje ia chover

Se hoje fez sol, preferimos P1e P2.

Na prática, problemas de incerteza envolvem probabilidade

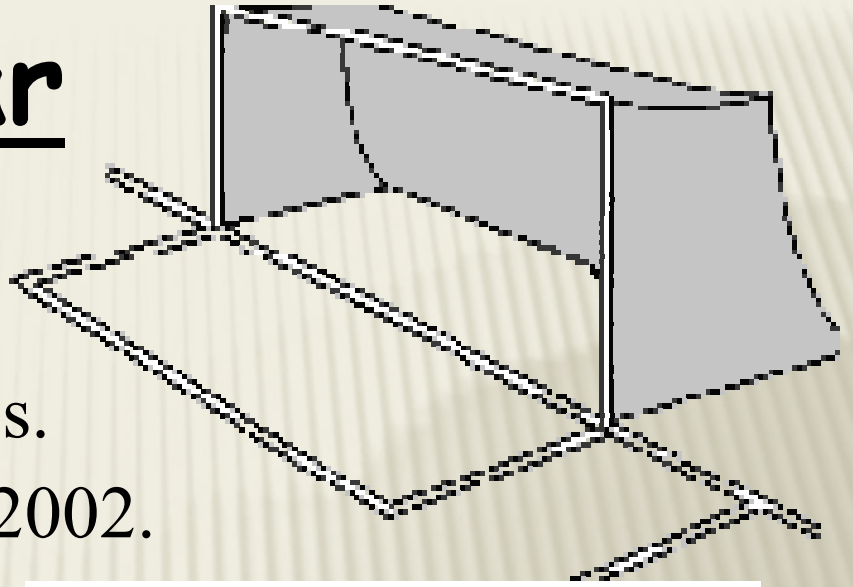
Para P1: P( sol ) = 80%
Para P2: P( sol ) = 70%
Para P3: P( sol ) = 40%

P1 e P2 acertaram...
mas P1 acertou mais

Princípio da máxima verossimilhança:
o melhor é quem fornece maior probabilidade para o que efetivamente ocorreu.

Voltaremos a esse ponto mais à frente...

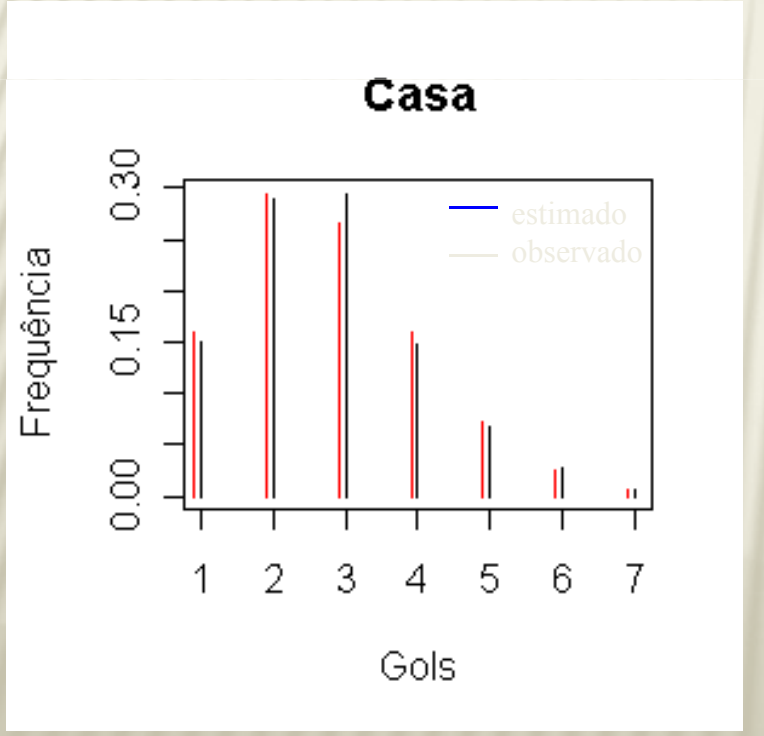
# Análise Preliminar

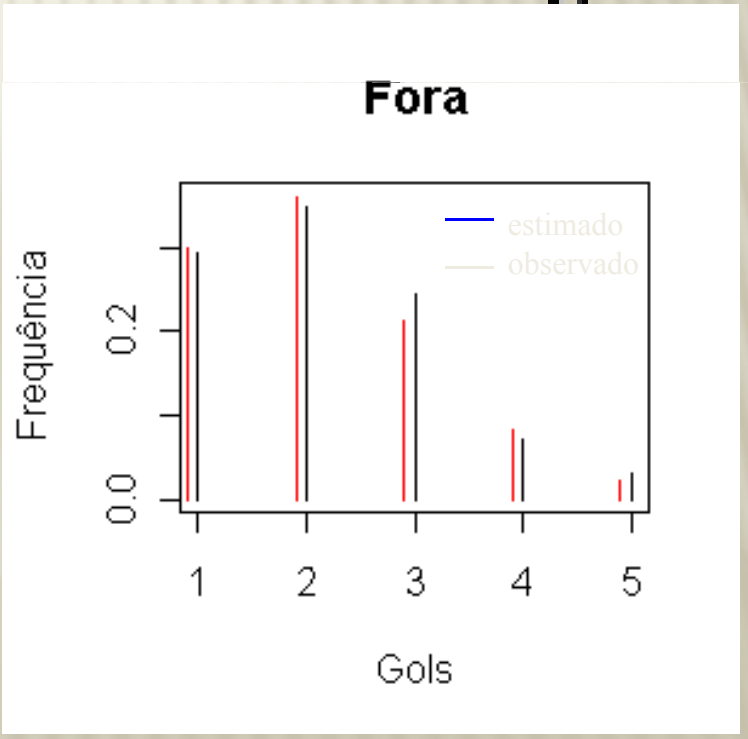


- Análise Univariada

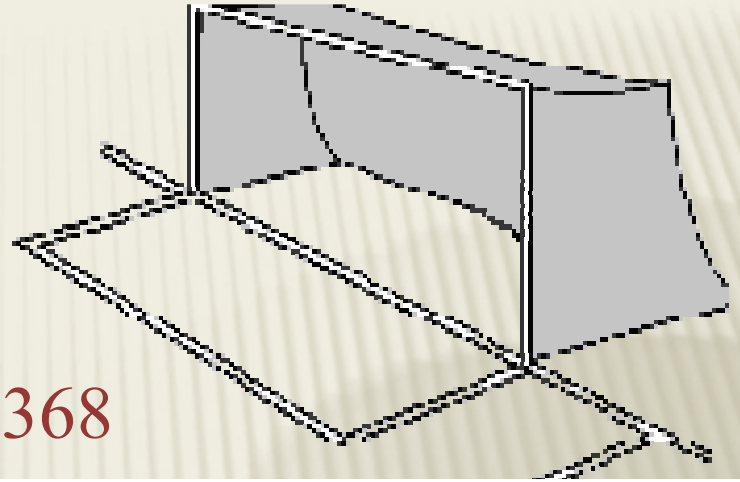
Poisson se ajusta bem aos dados.

Ex: Campeonato Brasileiro de 2002.

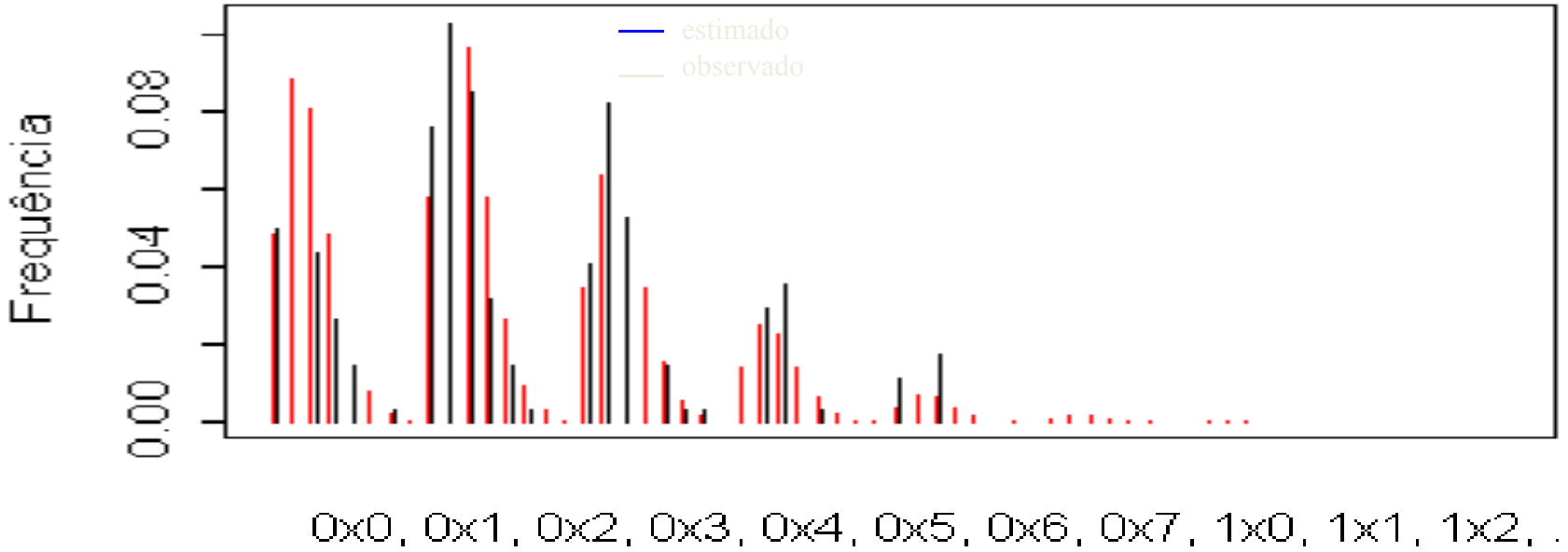


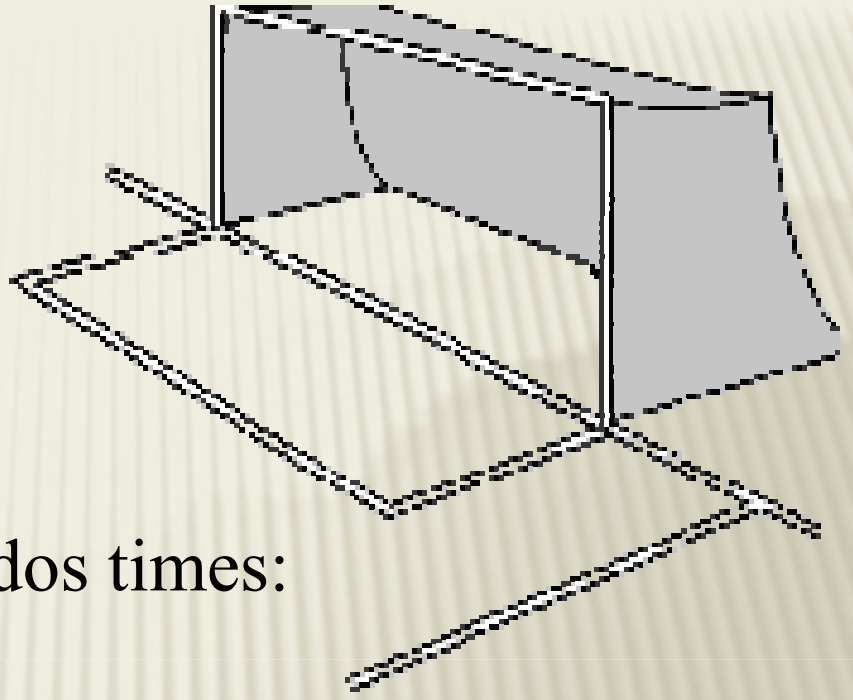
# Análise Preliminar



- Análise Bivariada
- $H_0$: Poisson Independentes
- Bondade de ajuste: p-valor = 0,368



# Modelo Inicial



Queremos explicar o resultado do jogo A x B.
Podemos postular fatores que determinam o comportamento dos times:

*Fator qualidade: quantifica o desempenho de um time; cada time tem seu fator qualidade*

*Fator Campo: informa o time que tem mando de campo; cada time tem o seu fator campo ou é um fator comum?*

Fator qualidade pode ser mais detalhado:

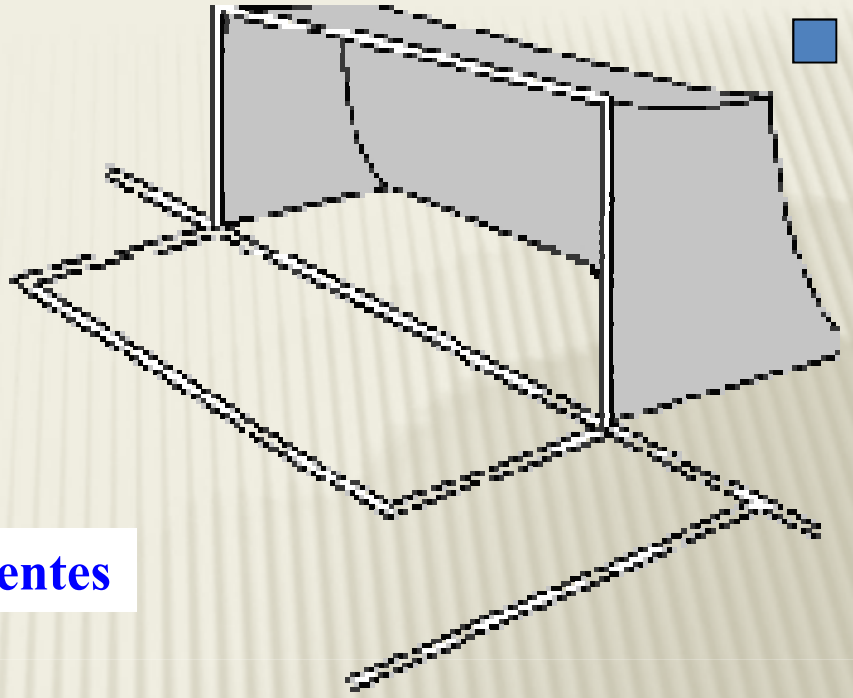
- pode ser fator único (força do time)
- pode ser decomposto em setores

Exemplos:

1. Fator ataque, Fator defesa, Fator meio de campo, ...
2. Fator infraestrutura, Fator elenco, ...

Vamos trabalhar com 2 fatores: ataque e defesa.

# Modelo Inicial



Assim, para o jogo A x B, temos o seguinte modelo:

$$NGF_A \sim Poisson(\lambda_A)$$
$$NGF_B \sim Poisson(\lambda_B)$$

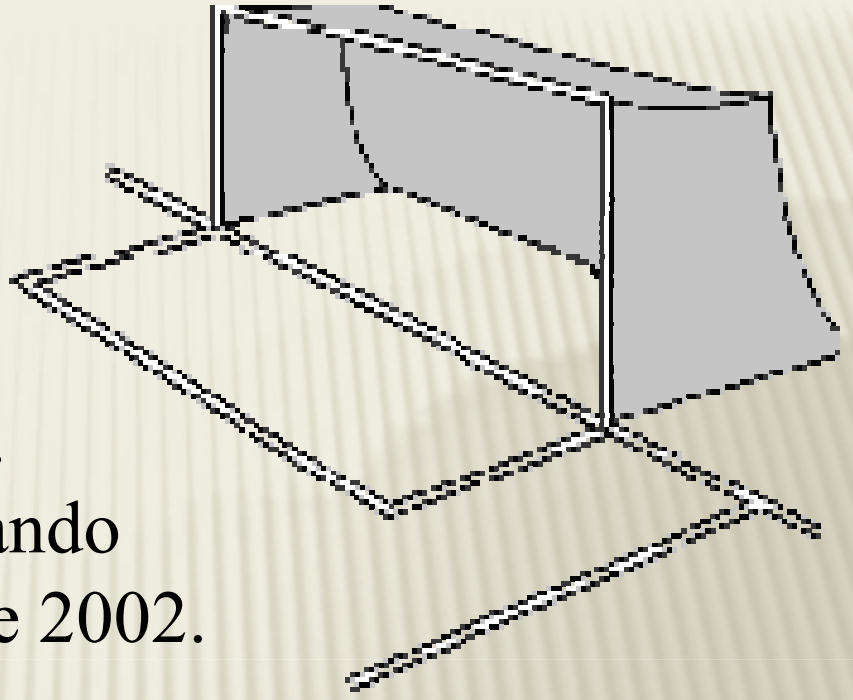
**Independentes**

$$\log \lambda_A = At_A - De_B + Ca_A$$
$$\log \lambda_B = At_B - De_A$$

onde:

| | |
|---|---|
| $NGF_{time}$ | representa o número de gols feitos pelo *time* |
| $At_{time}$ | representa o fator ataque do *time* |
| $De_{time}$ | representa o fator defesa do *time* |
| $Ca_{time}$ | representa o fator campo do *time* |

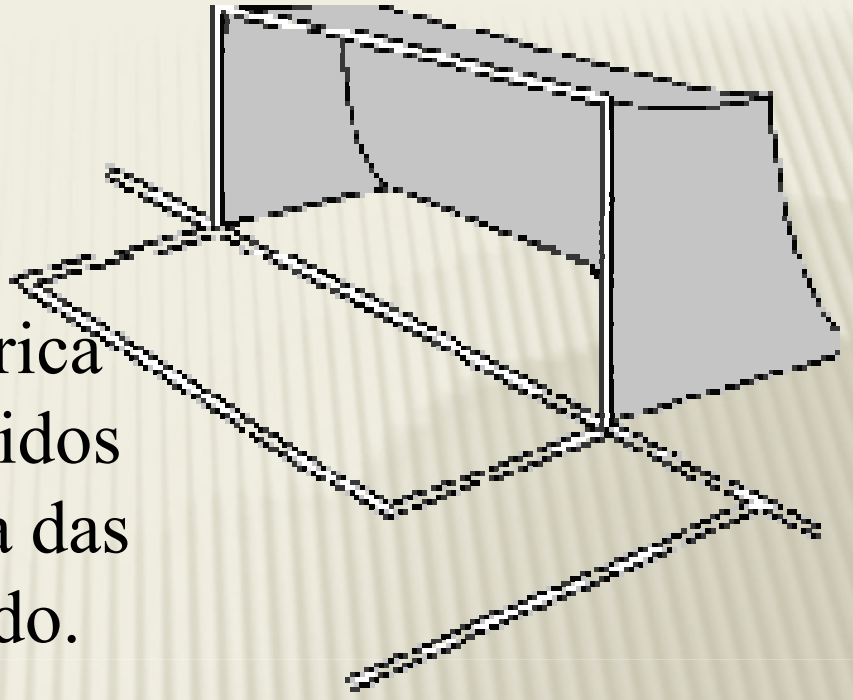
# Modelo Inicial



Abaixo, temos a tabela com os fatores para os times do Rio. Esses fatores foram obtidos usando primeira fase do campeonato de 2002.

| | Fator Ataque | Fator Defesa | Fator Campo | Gols Pró | Gols Contra |
|---|---|---|---|---|---|
| Botafogo | -0.873 | -0.063 | 0.264 | 24 | 39 |
| Flamengo | -0.451 | -0.005 | 0.346 | 38 | 39 |
| Fluminense | -0.416 | 0.080 | 0.473 | 43 | 46 |
| Vasco | -0.363 | -0.172 | 0.122 | 37 | 38 |

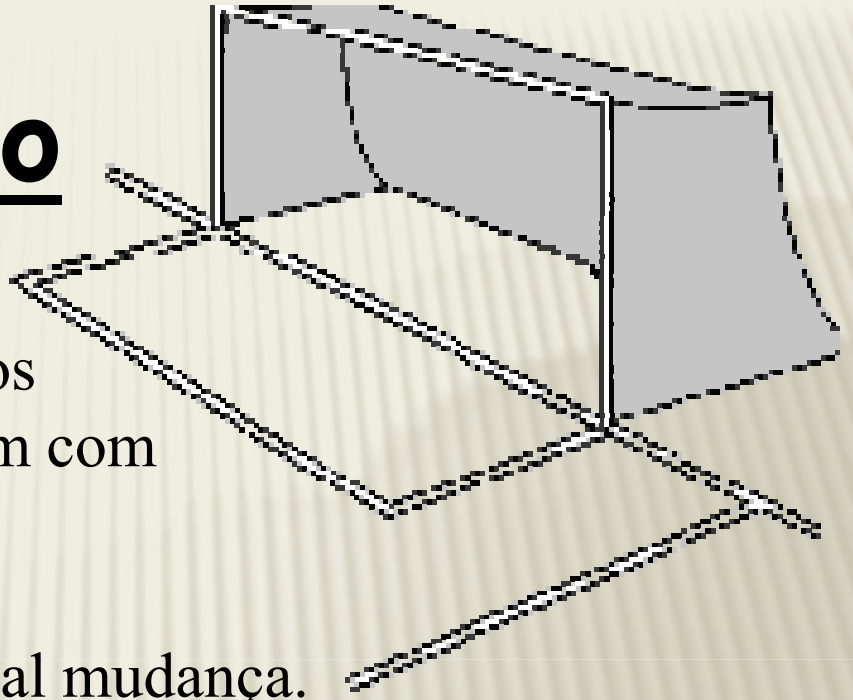
# Modelo Inicial



Agora, com 3 seleções da América do Sul. Esses fatores foram obtidos usando os dados até a 7ª rodada das Eliminatórias da Copa do Mundo.

| | Fator Ataque | Fator Defesa | Fator Campo | Gols Pró | Gols Contra |
|---|---|---|---|---|---|
| Brasil | -0.62 | -0.33 | 0.31 | 11 | 7 |
| Equador | -1.70 | -0.03 | 1.32 | 8 | 7 |
| Uruguai | -0.27 | 0.90 | 0.04 | 12 | 19 |

# Modelo Dinâmico



Estávamos supondo até agora que os parâmetros do modelo não variavam com as rodadas.

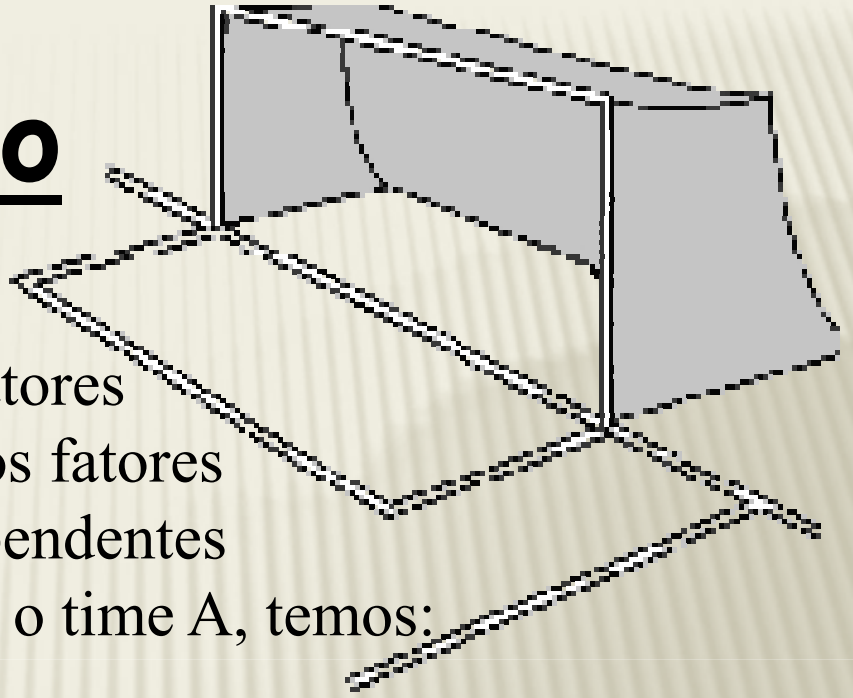
Agora, achamos razoável permitir tal mudança.

Portanto, $At_{time}$ virou vetor.

Ou seja, temos agora: $At^{1}_{time}\, At^{2}_{time}\, , \ldots, At^{T}_{time}$.

onde T é o número total de rodadas

# Modelo Dinâmico



Achamos razoável assumir que os fatores na rodada $i+1$ dependem dos mesmos fatores na rodada $i$, ou seja, são sempre dependentes do passo anterior. Por exemplo, para o time A, temos:

Fator Ataque

$$At_A^{i+1} = At_A^i + \omega_{At}^{i+1} \quad \text{onde } \omega_{At}^{i+1} \sim N(0, \sigma^2_{At})$$

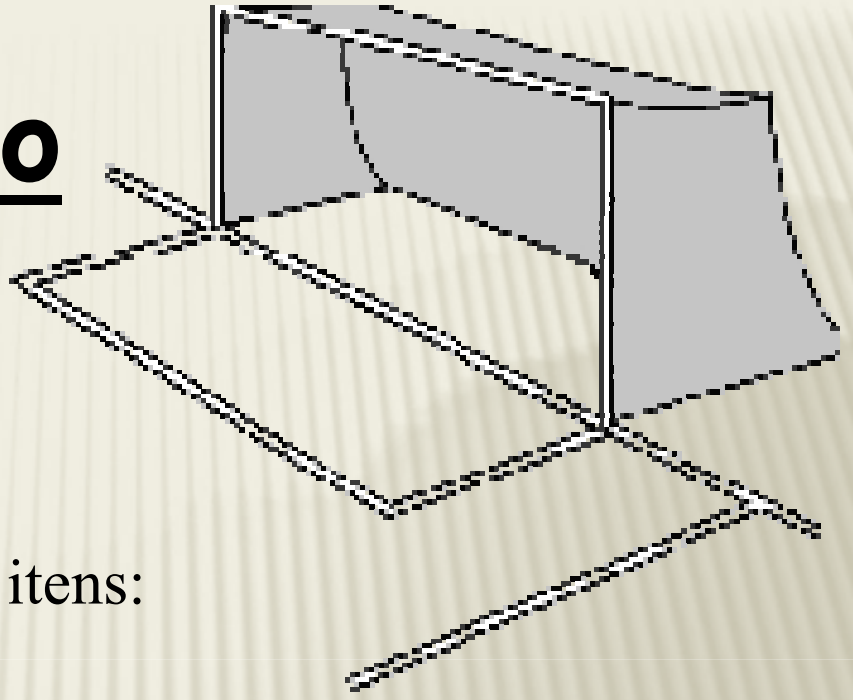
Fator Defesa

$$De_A^{i+1} = De_A^i + \omega_{De}^{i+1} \quad \text{onde } \omega_{De}^{i+1} \sim N(0, \sigma^2_{De})$$

Fator Campo

$$Ca_A^{i+1} = Ca_A^i + \omega_{Ca}^{i+1} \quad \text{onde } \omega_{Ca}^{i+1} \sim N(0, \sigma^2_{Ca})$$

# Modelo Dinâmico



O modelo é completado com mais 2 itens:

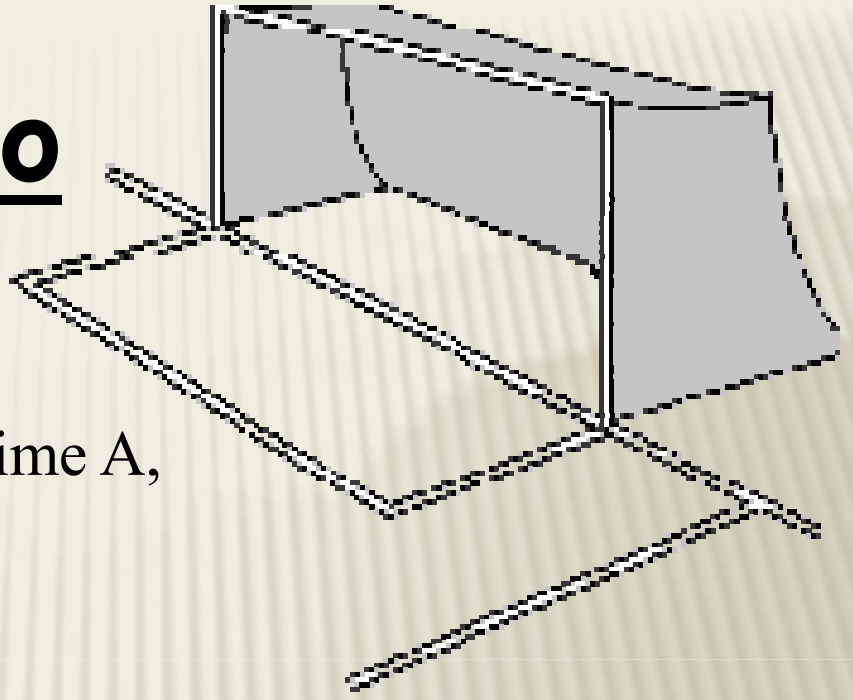
- as volatilidades $\sigma^2_{At}$, $\sigma^2_{De}$ e $\sigma^2_{Ca}$ das perturbações $\omega_{At}^{i}$, $\omega_{De}^{i}$ e $\omega_{Ca}^{i}$ são obtidas empiricamente.
- a priori para os parâmetros da rodada inicial para os times. pode ser baseada no desempenho passado ou ser vaga:

$$At^{1}_{time} \sim N(0, 10^4)$$

$$De^{1}_{time} \sim N(0, 10^4)$$

$$Ca^{1}_{time} \sim N(0, 10^4)$$

# Modelo Dinâmico



Considere o jogo A x B

O modelo para as observações do time A, jogando em casa, agora é

$$NFG_A^i \sim Poisson\left(\lambda_A^i\right)$$

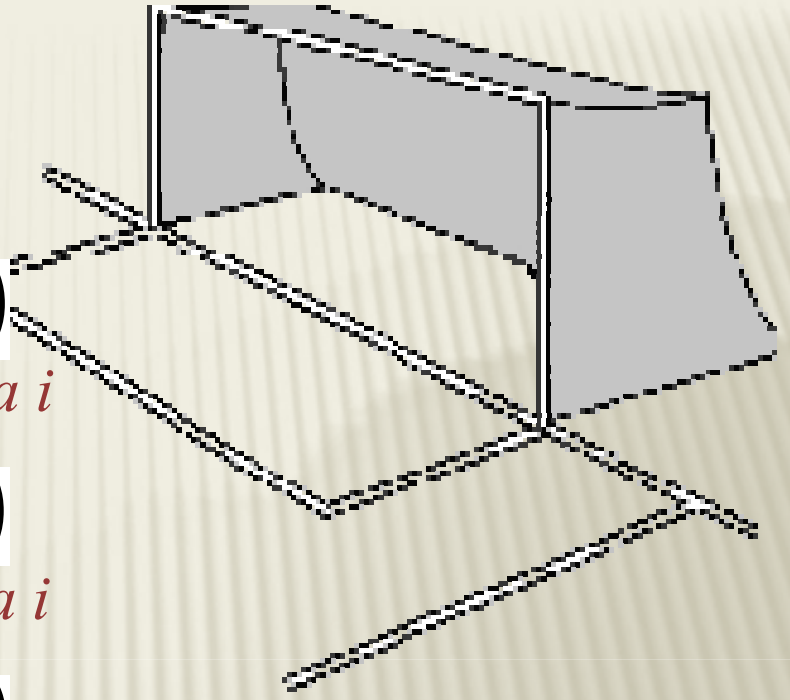
$$\log \lambda_A^i = At_A^i - De_B^i + Ca_A^i$$

Da mesma forma, para o time B, temos:

$$NFG_B^i \sim Poisson\left(\lambda_B^i\right)$$

$$\log \lambda_B^i = At_B^i - De_A^i$$

# Notação



$$At^i = \left(At^i_{Atletico-MG}, At^i_{Atletico-PR}, \ldots, At^i_{Vitoria}\right)$$

vetor com fatores ataque para a *rodada i*

$$De^i = \left(De^i_{Atletico-MG}, De^i_{Atletico-PR}, \ldots, De^i_{Vitoria}\right)$$

vetor com fatores defesa para a *rodada i*

$$Ca^i = \left(Ca^i_{Atletico-MG}, Ca^i_{Atletico-PR}, \ldots, Ca^i_{Vitoria}\right)$$

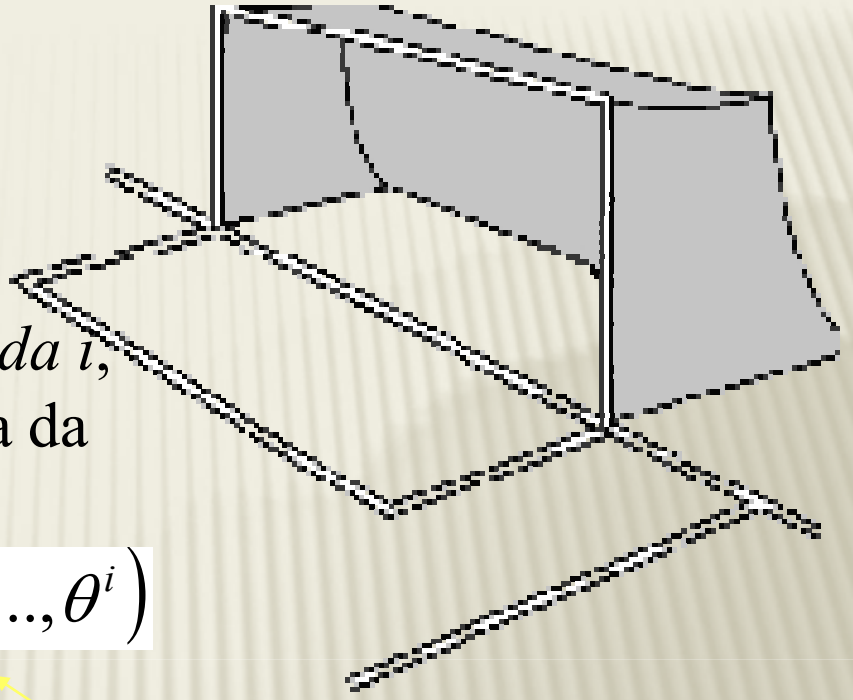
vetor com fatores campo para a *rodada i*

$\theta^i = \left(At^i, De^i, Ca^i\right)$ vetor de parâmetros para a *rodada i*

$NGF^i = (NGF^i_{AtleticoMG}, \ldots, NGF^i_{Vitoria})$ número de gols feitos na *rodada i*

$D^i = \{NGF^1, \ldots, NGF^i\}$ todas as informações até a *rodada i*

# Estimação



Utilizando o teorema de Bayes, a estimação dos parâmetros até a *rodada i*, será feita a partir da posteriori obtida da seguinte forma:

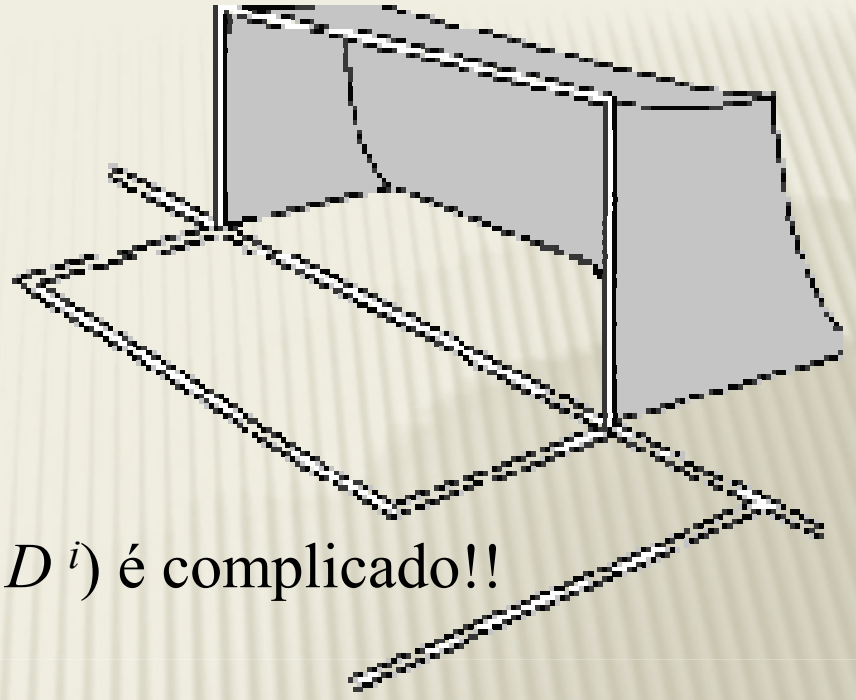
$$p\left(\theta^1,...,\theta^i \mid D^i\right) \propto L\left(\theta^1,...,\theta^i\right) p\left(\theta^1,...,\theta^i\right)$$

posteriori verossimilhança priori

verossimilhança: $L\left(\theta^1,...,\theta^i\right) = \prod_{t=1}^{i} L\left(\theta^t\right)$ e $L\left(\theta^t\right) = \prod_{j=AtleticoMG}^{Vitoria} p\left(NGF_j^t \mid \theta^t\right)$

priori: $p\left(\theta^1,...,\theta^i\right) = \prod_{t=2}^{i} p\left(\theta^t \mid \theta^{t-1}\right) p\left(\theta^1\right)$

# Computação



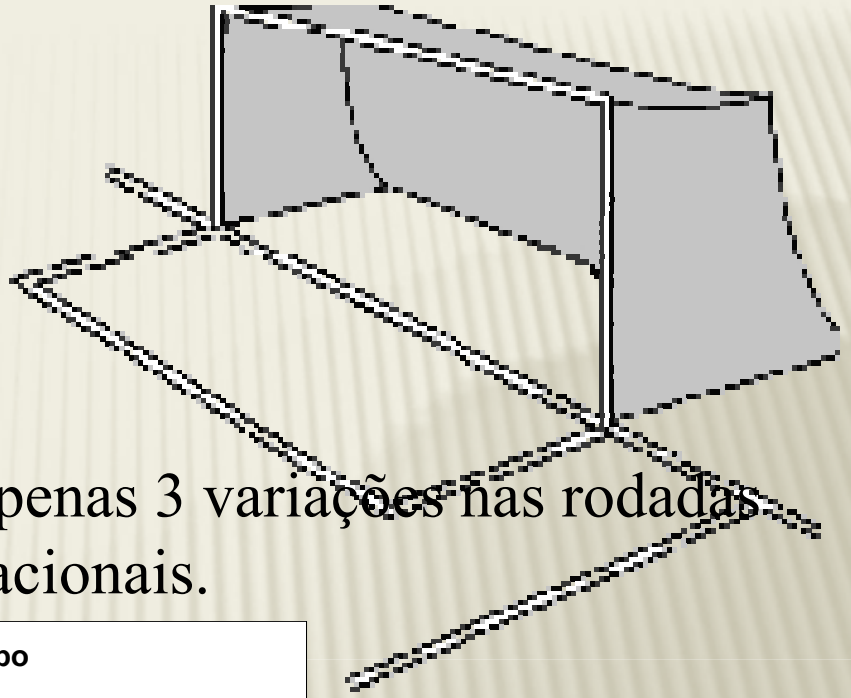
Extrair informações de p($\theta^1$,..., $\theta^i \mid D^i$) é complicado!!

Esse problema é solucionado através de simulações via MCMC (*Gamerman e Lopes, 2006*). Um programa utilizado para fazer tais simulações é o WinBugs (*Spiegelhalter et al, 2003*).

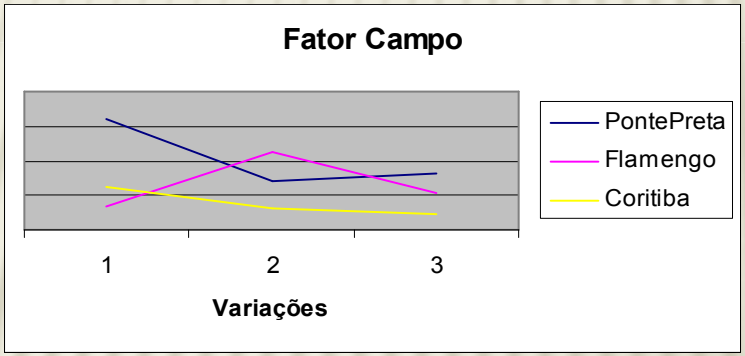
Dessa forma, serão obtidas amostras da posteriori.

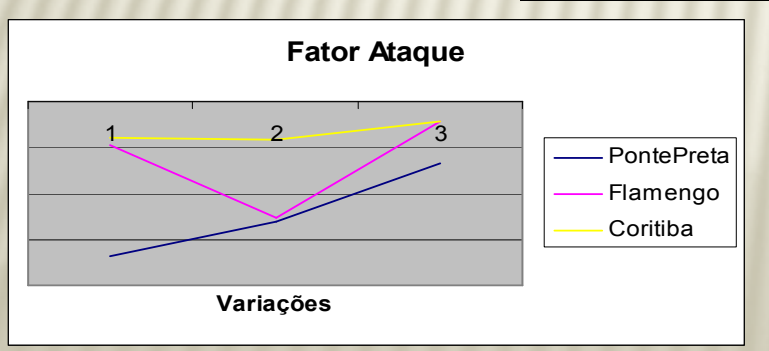
E portanto, teremos amostras de $\theta^i \mid D^i$, para determinada *rodada i*.

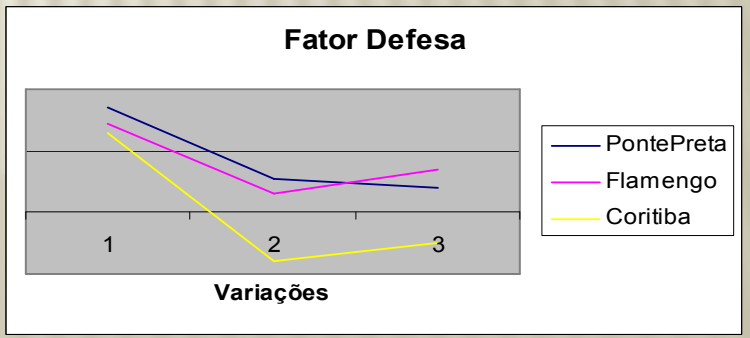
# Computação



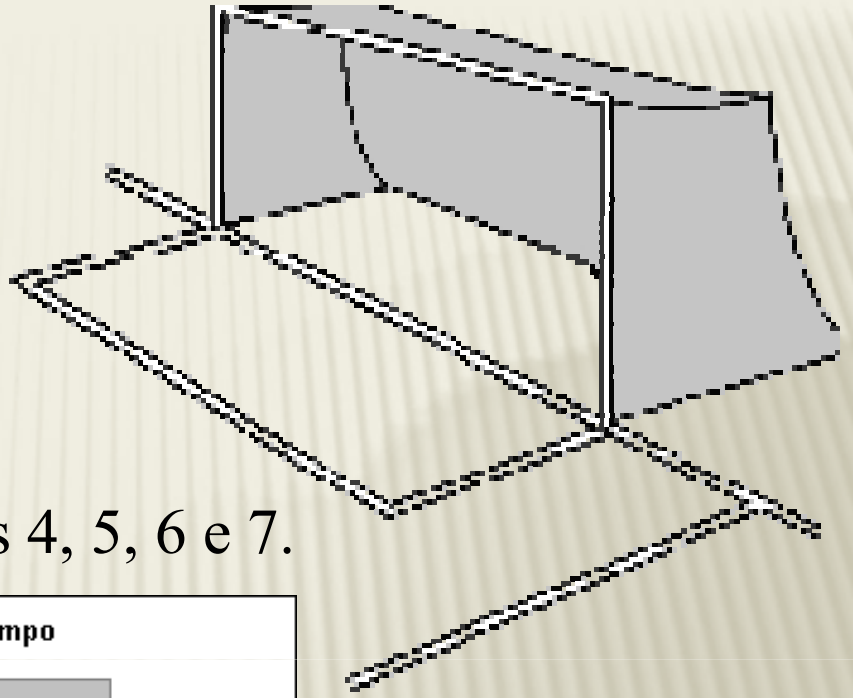
Exemplo: Camp. Brasileiro de 2002
parâmetros de 3 times:
Coritiba, Flamengo e Ponte Preta. Apenas 3 variações nas rodadas 15, 30 e 44 devido a limites computacionais.



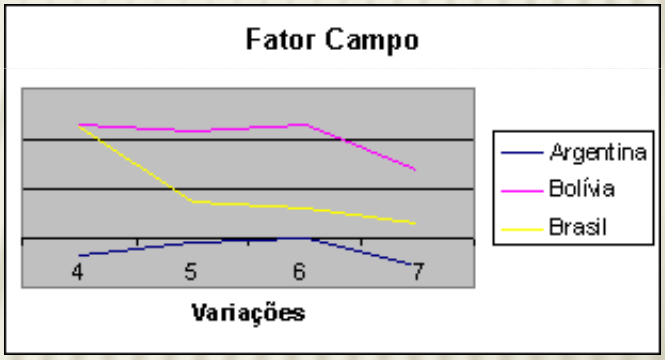



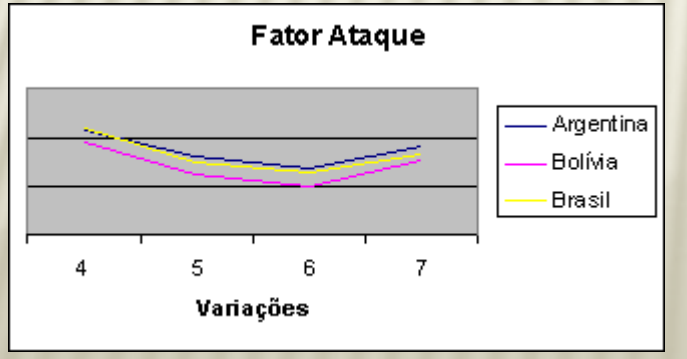



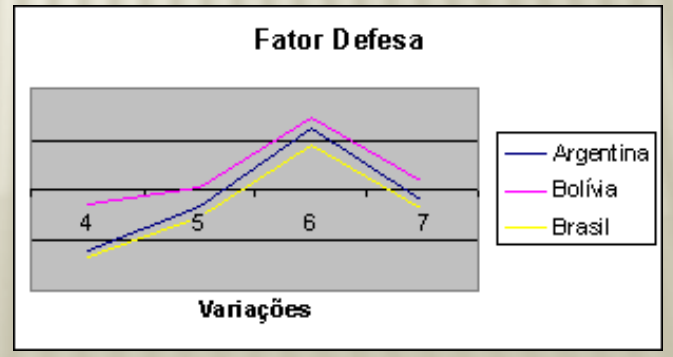
# Computação



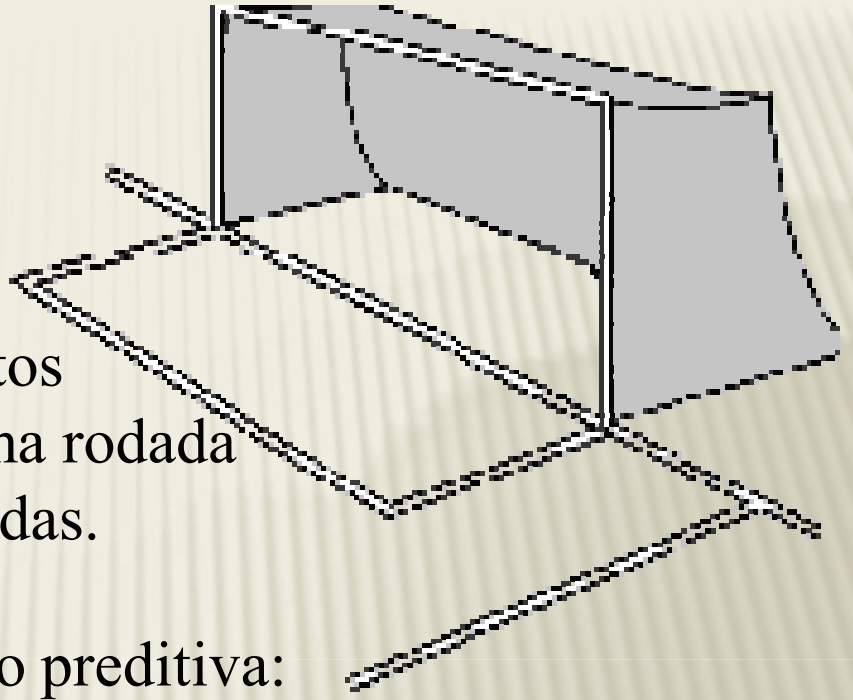
Outro exemplo: Copa do Mundo
parâmetros de 3 países:
Argentina, Bolívia e Brasil.
Foram feitas 4 variações nas rodadas 4, 5, 6 e 7.



# Previsões



Aqui, vamos obter os valores previstos para o número de gols feitos para uma rodada futura, a partir de informações passadas.

A previsão é baseada na distribuição preditiva:

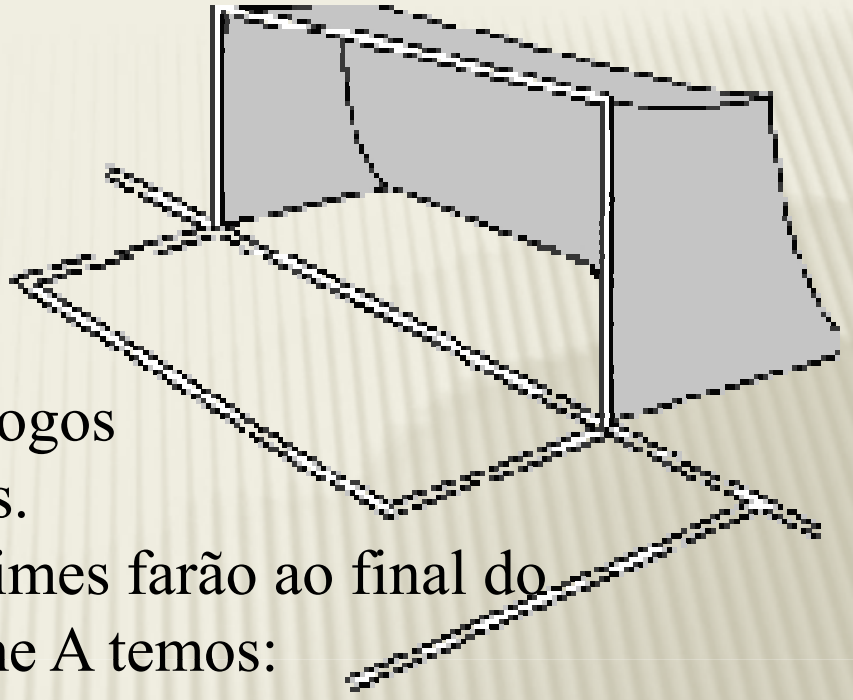
$$p\left(NGF^{i+h} \mid D^i\right) = \int p\left(NGF^{i+h} \mid \theta^i, D^i\right) p\left(\theta^i \mid D^i\right) d\theta^i$$

1 2 3

onde: $NGF^{i+h} \mid \theta^i, D^i \sim Poisson\left(\lambda^{i+h}\right)$

3 é obtido por simulação via MCMC, servindo de parâmetro para simular amostras de 2. Desta forma, automaticamente temos amostras de 1.

# Previsões



Com as distribuições preditivas dos jogos podemos calcular várias distribuições.
Exemplo: número de pontos que os times farão ao final do campeonato. Por exemplo, para o time A temos:

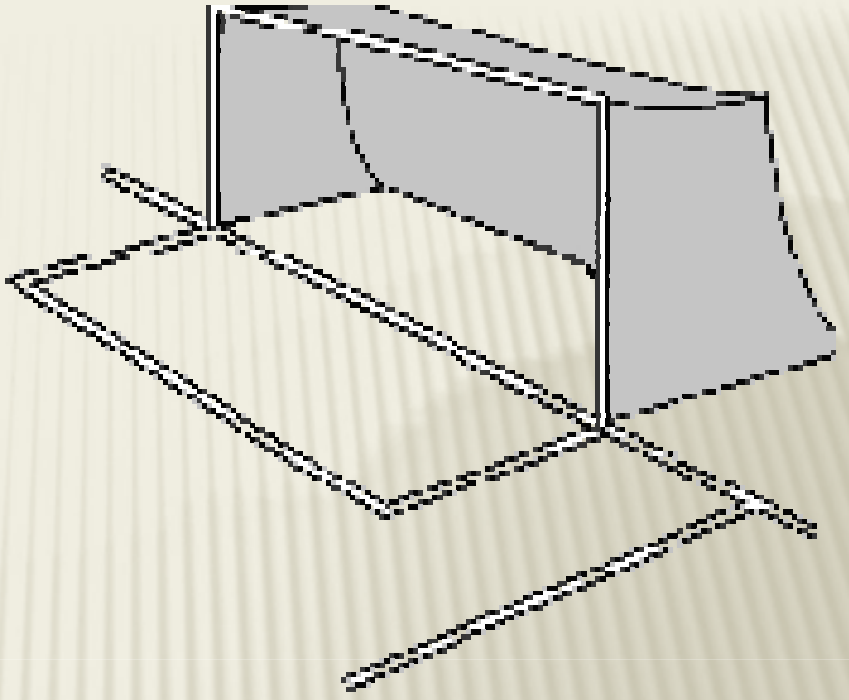
$$NP_A^T = f\left(NGF^1,\ldots,NGF^T\right)$$

$NP^T_A$ é o número de pontos do *time A* na rodada final *T*

Qualquer função desse tipo pode ter
sua distribuição aproximada por simulação

Exemplo: classificação (que depende não só de NP).

# Resultados



Aqui, é possível calcular as probabilidades para o resultado de cada jogo (1x0, 2x0, ...).

Para exemplificar, será exposto um resultado mais detalhadamente.

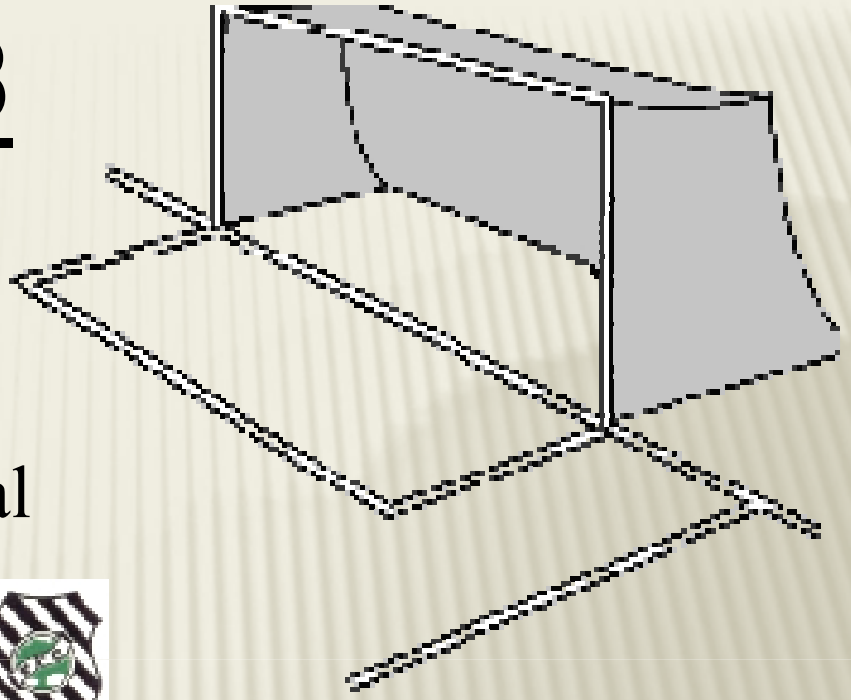
# Resultados 2003

Vitória
1x0 15.2%
2x0 9.7%
2x1 8.9%
3x0 4.0%
3x1 3.3%
3x2 1.5%
Outros 3.6%

Empate
0x0 9.8%
1x1 14.4%
2x2 3.6%
3x3 0.3%
Outros 0.1%

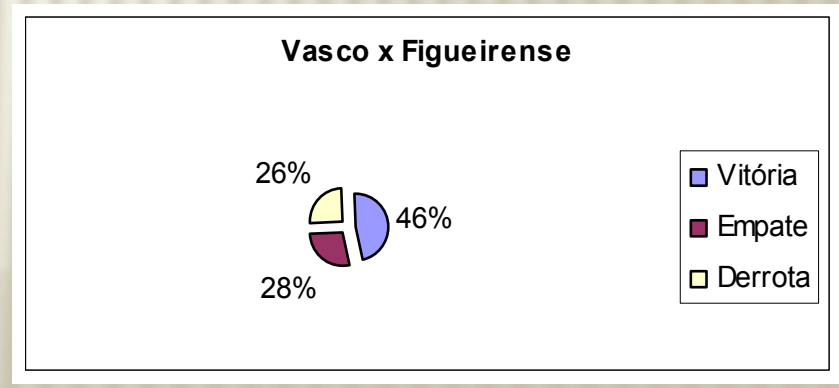
Derrota
0x1 10.8%
0x2 3.6%
1x2 5.5%
0x3 1.3%
1x3 1.9%
2x3 1.0%
Outros 1.5%

**Os 2 resultados mais prováveis**



resultado real

**1 x 0**



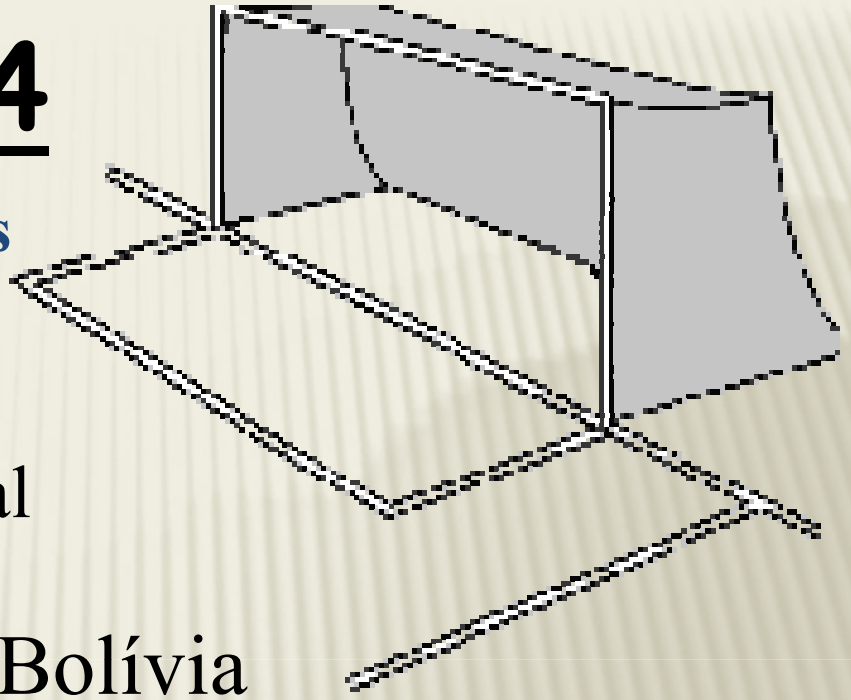
# Resultados 2004

Vitória
1x0 9.7%
2x0 15.7%
2x1 8.6%
3x0 19.9%
3x1 14.1%
3x2 2.0%
4x0 11.9%
4x1 5.2%
Outros 0.9%

**Os 3 resultados mais prováveis**

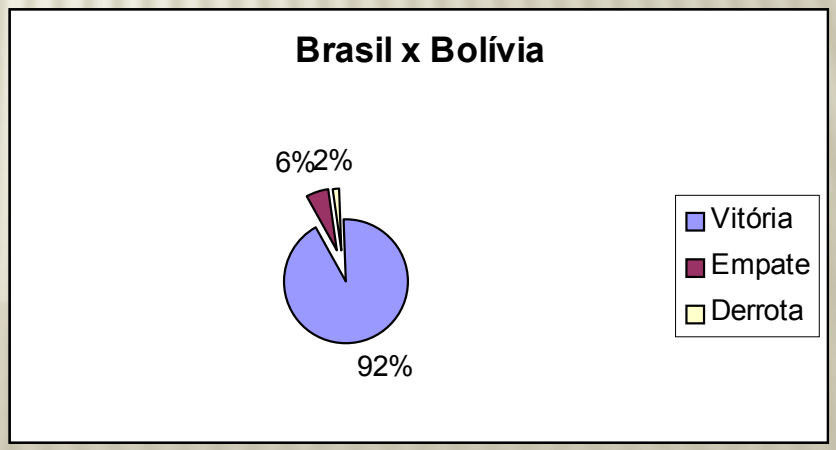
Empate
0x0 2.0%
1x1 2.5%
2x2 1.3%
3x3 0.1%
Outros 0.1%

Derrota
0x1 0.7%
0x2 0.1%
1x2 0.8%
0x3 0.1%
1x3 0.1%
2x3 0.1%
Outros 0.1%

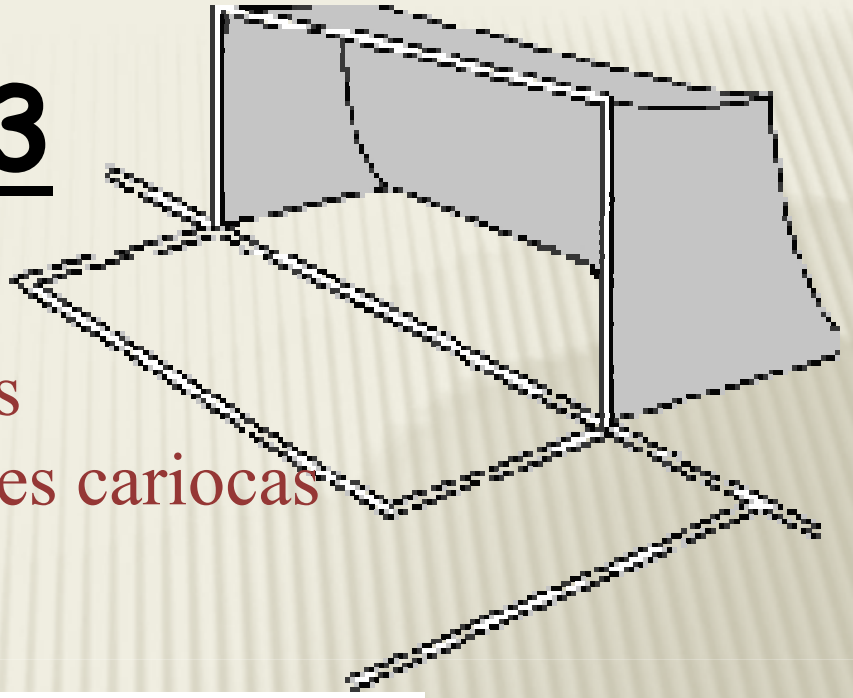


resultado real

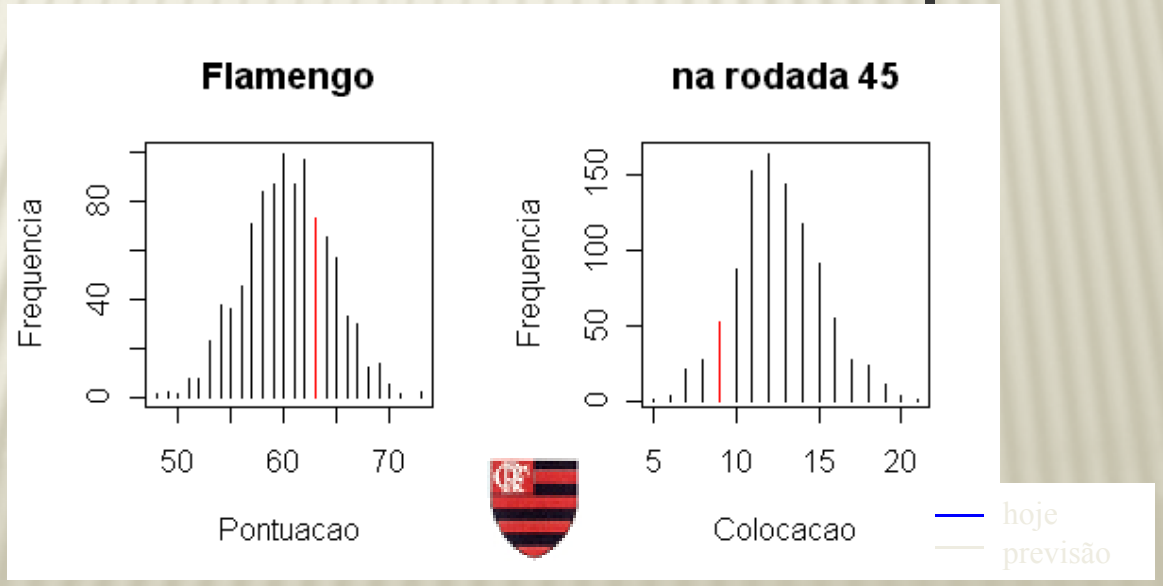
Brasil ? x ? Bolívia



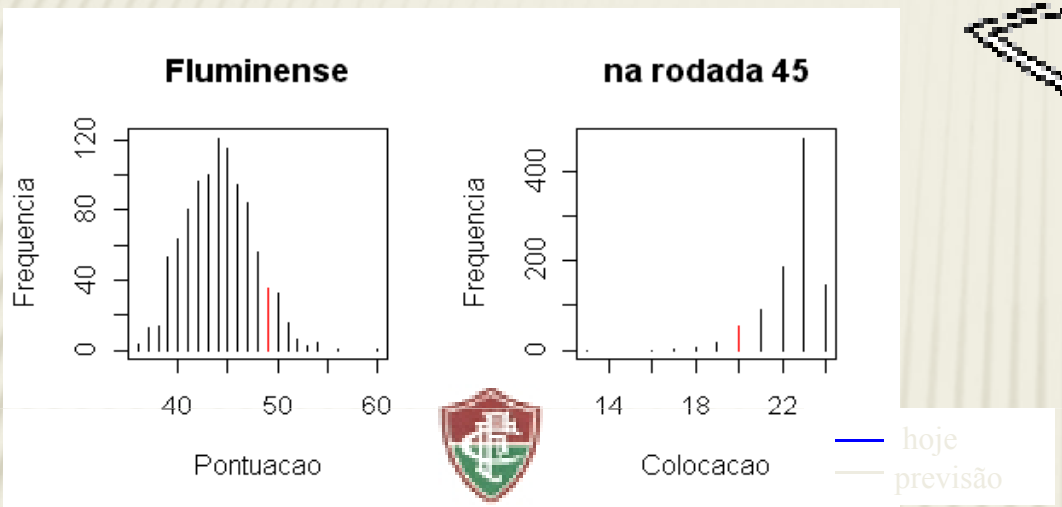
# Resultados 2003

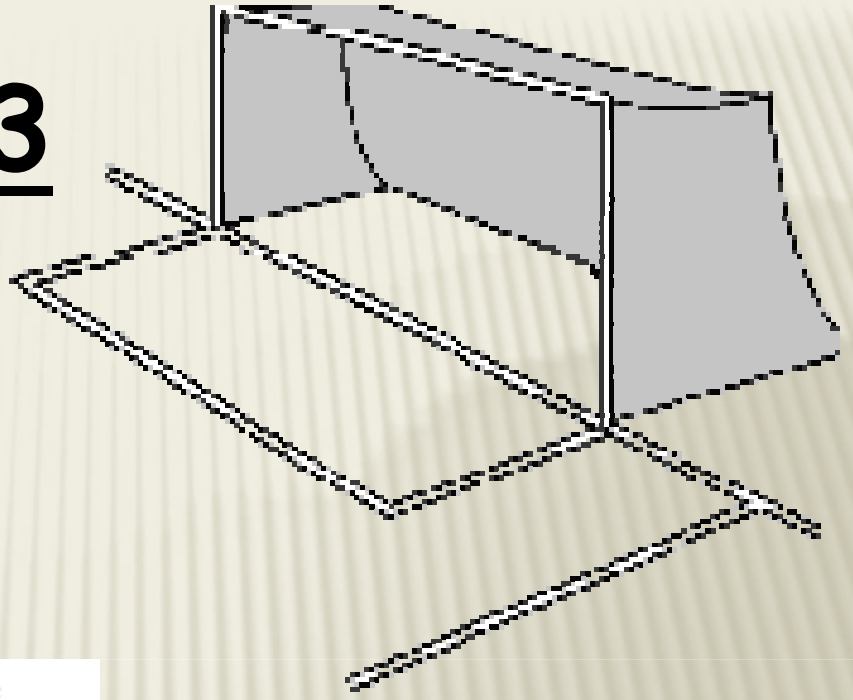


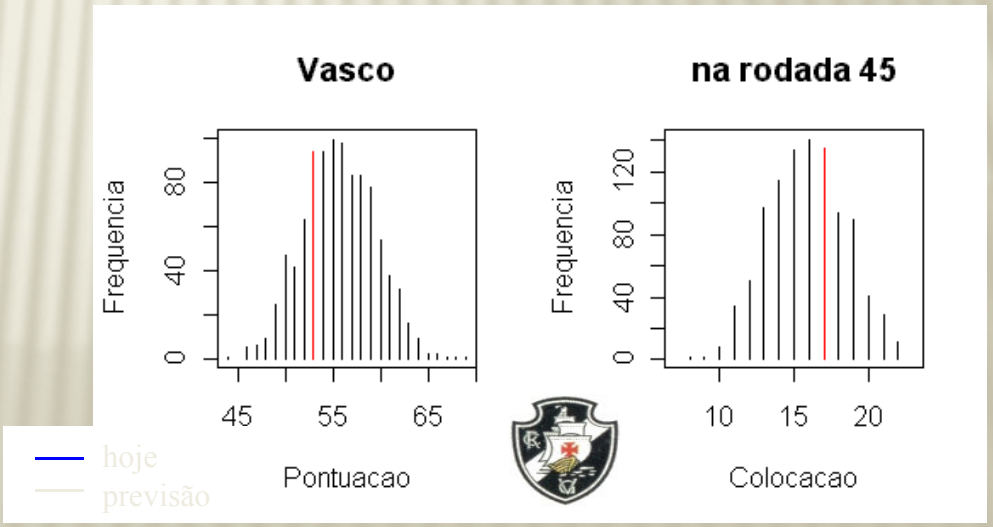
Na rodada de número 34, foi feita uma análise e chegamos às seguintes previsões para os times cariocas na rodada 45:



# Resultados 2003

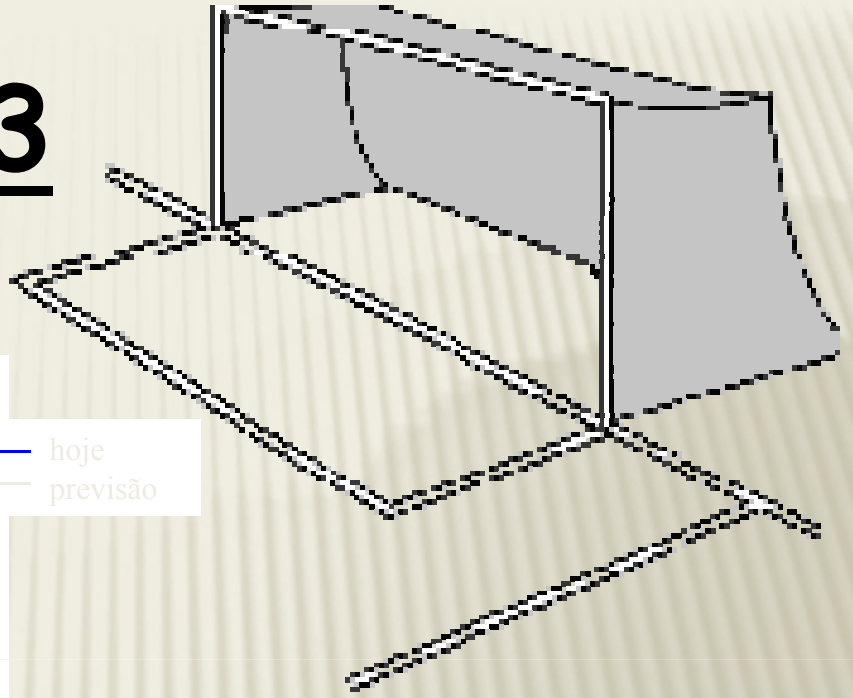


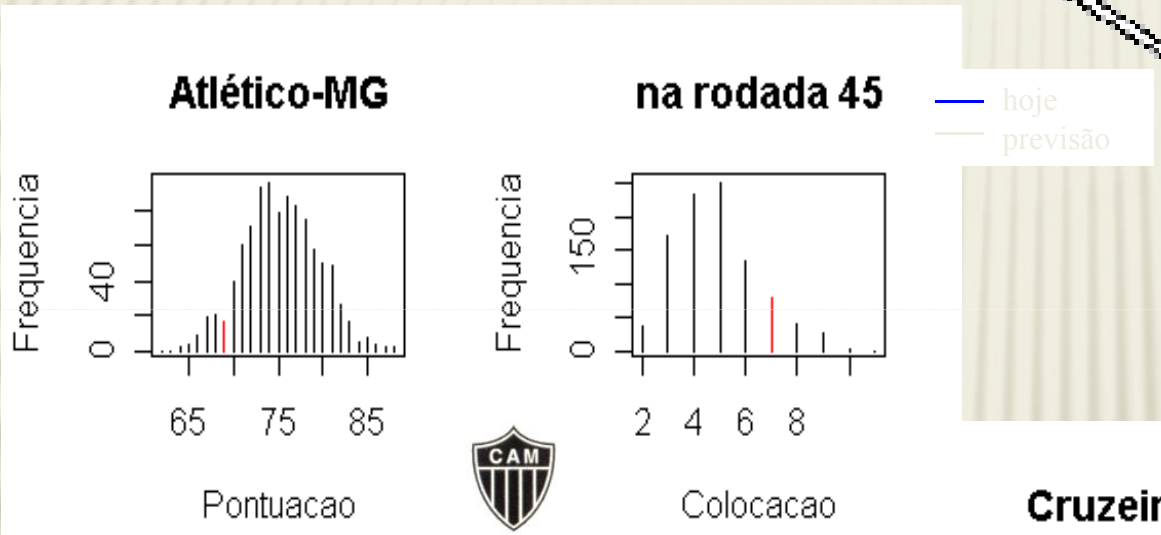



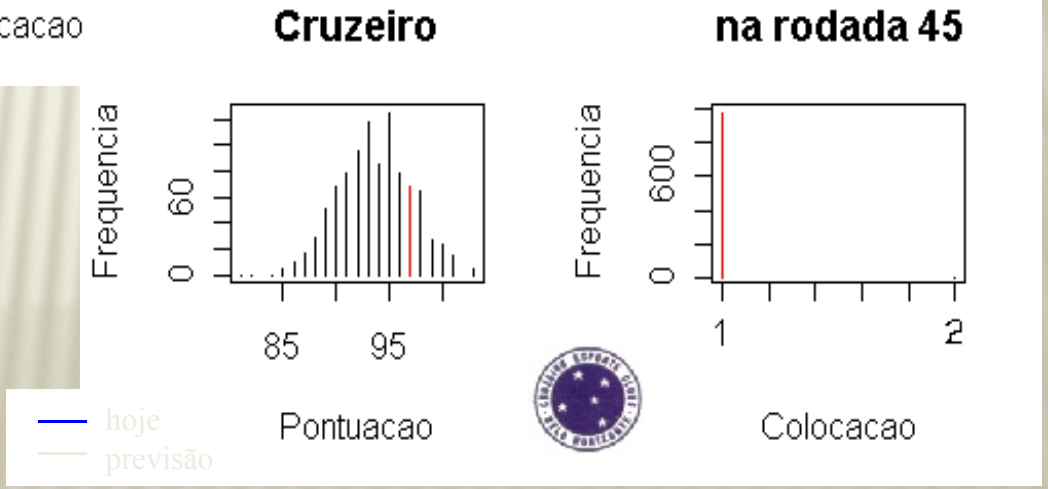


# Resultados 2003

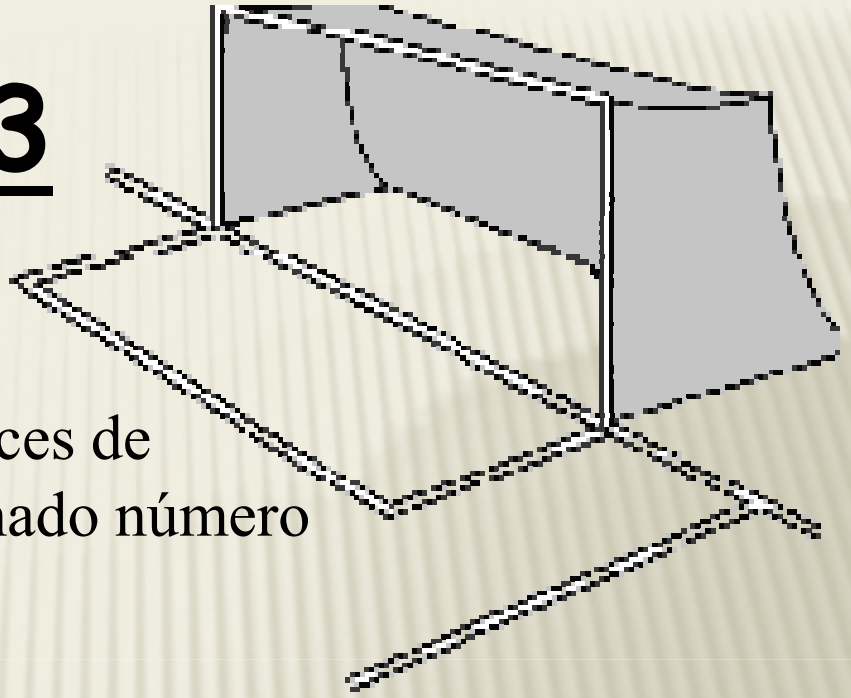
para os times mineiros, temos:





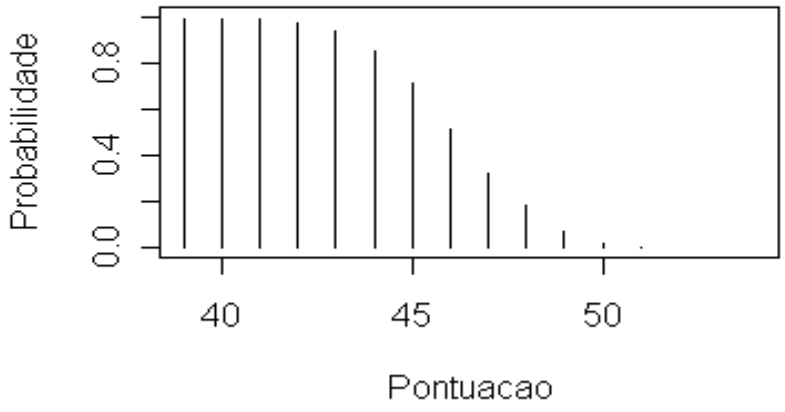



# Resultados 2003

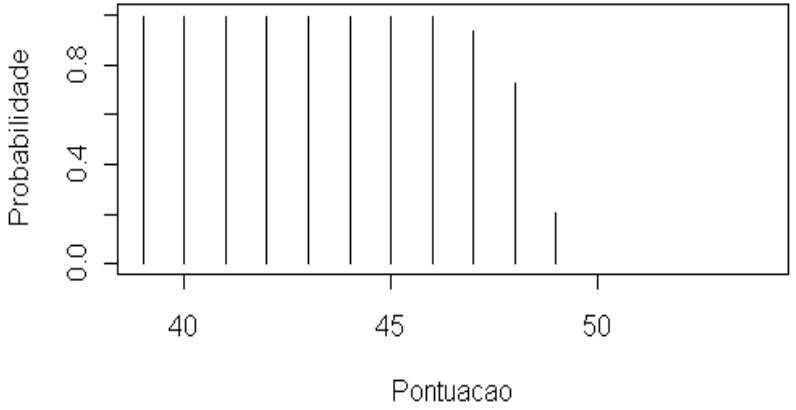


Os gráficos abaixo mostram as chances de um time ser rebaixado com determinado número de pontos em duas rodadas distintas.

## Rodada 34



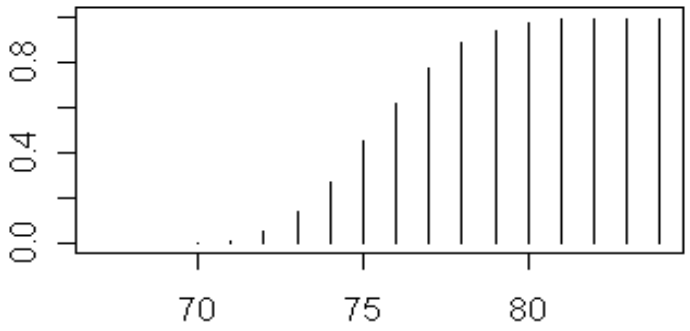

## Rodada 45



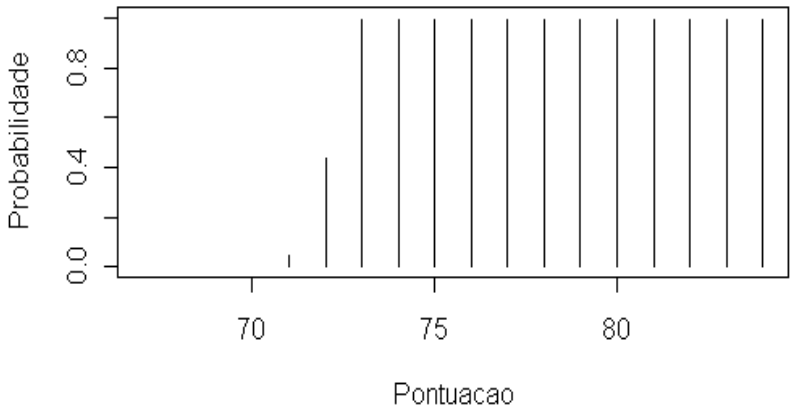
# Resultados 2003

Os gráficos abaixo mostram as chances de um time se classificar para a Libertadores com determinado número de pontos em duas rodadas distintas.

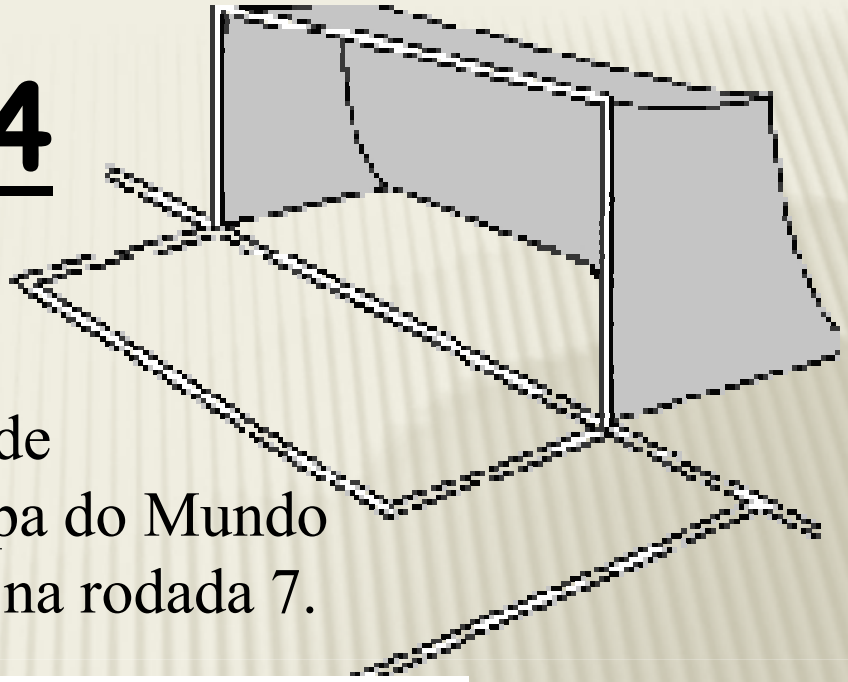
Rodada 34



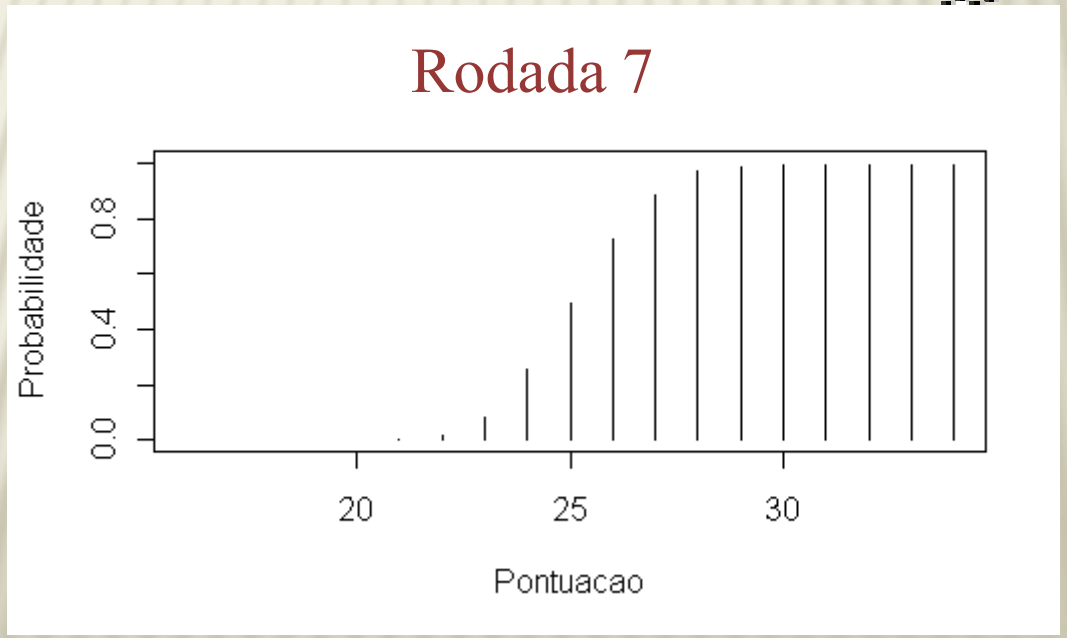

Rodada 45



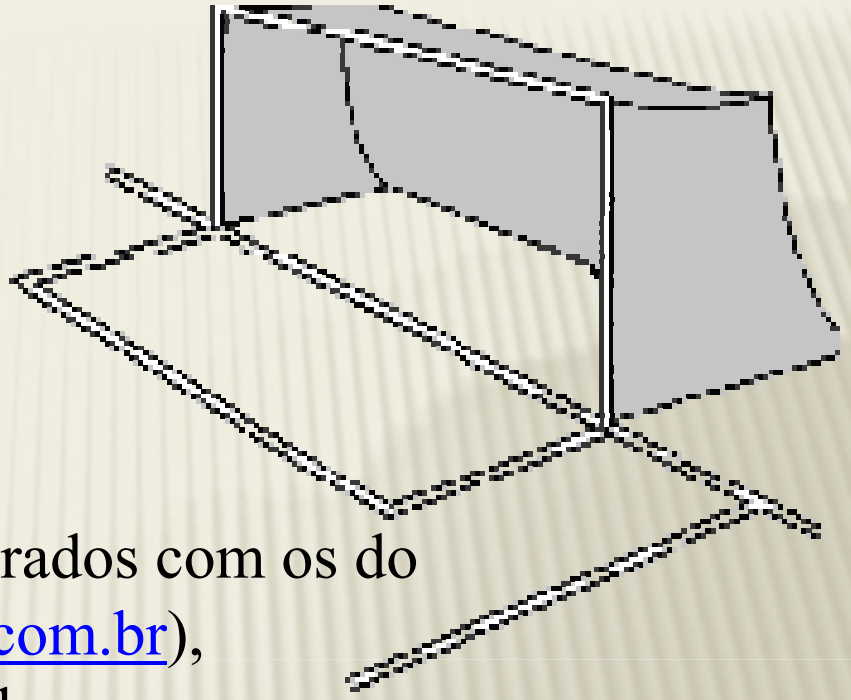
# Resultados 2004



O gráfico abaixo mostra as chances de uma seleção se classificar para a Copa do Mundo com determinado número de pontos na rodada 7.



# Análise de Resultados



Resultados do nosso modelo comparados com os do Chance de Gol (www.chancedegol.com.br), utilizandos o critério das verossimilhanças

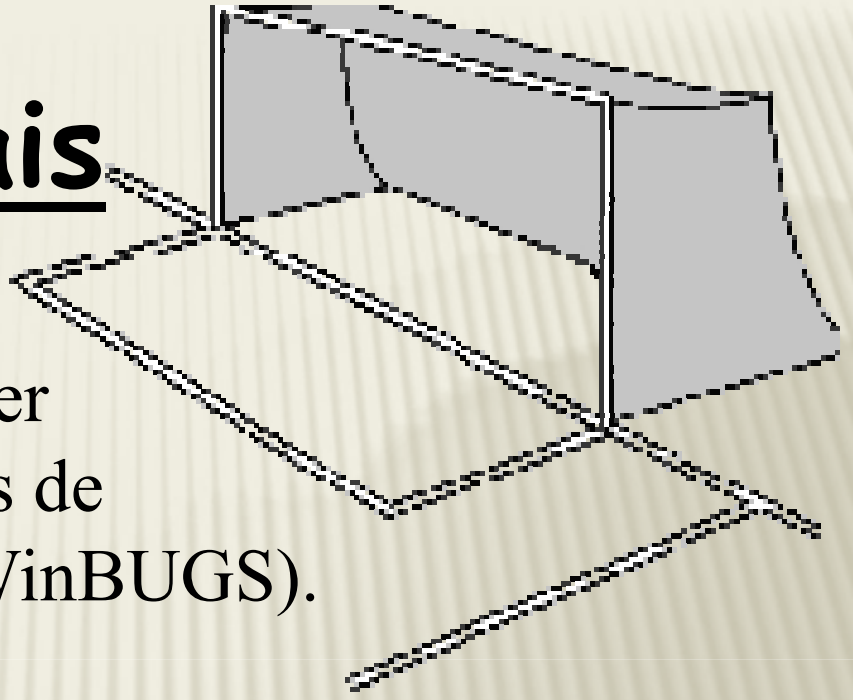
$$Verossimilhança = P(EO_1, ..., EO_T)$$

$EO_i$ é o Evento Ocorrido no jogo i

Verossimilhança do modelo do Chance de Gol: $2.26 \times 10^{-17}$

Verossimilhança do nosso modelo: $7.66 \times 10^{-17}$

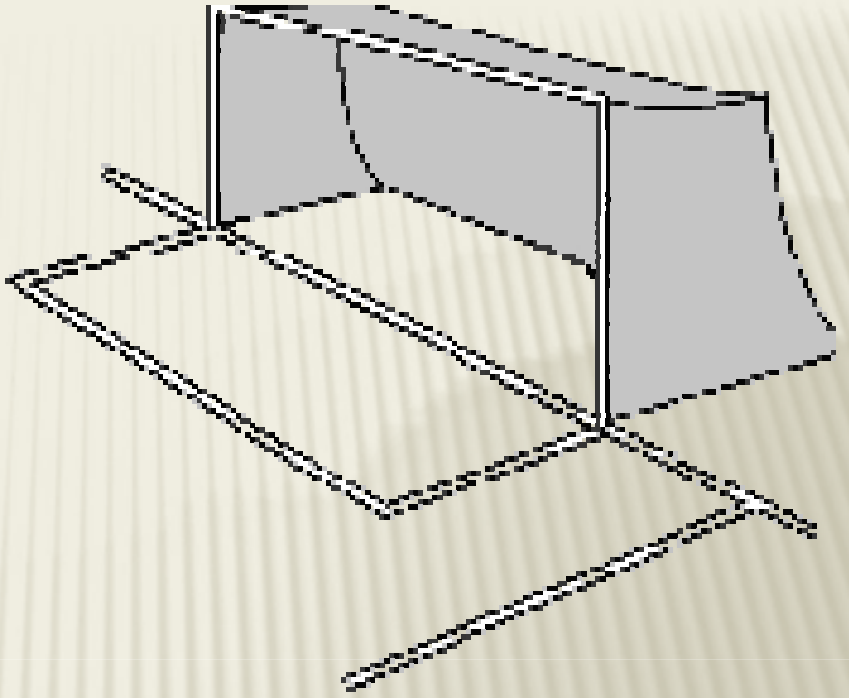
# Comentários finais



- Modelos válidos em qualquer campeonato e muito simples de serem implementados (no WinBUGS).
- Modelo dinâmico é mais razoável.
- Modelo pode ser estendido/alterado em várias direções.
- Dissertação de Fabio F. Farias (2008) apresenta extensões melhoradoras ao permitir evoluções estacionárias para os fatores.

# Bibliografia


- Farias, F. F. (2008). Análise e previsão de resultados de partidas de futebol. Dissertação de mestrado, Estatística, UFRJ.
- Gamerman, D. e Lopes, H. (2006) Markov Chain Monte Carlo. 2ª. Edição. Nova York: Chapman & Hall.
- Knorr-Held, L. (2000) Dynamic rating of sports teams. The Statistician (JRSS-D), 49, 261-276.
- Rue, H. e Salvesen O. (2000) Prediction and retrospective analysis of soccer matches in a league. JRSS-D, 49, 399-418.
- Spiegelhalter, D., Thomas, A., Best, N. e Lunn, D. (2003) WinBugs User Manual. Cambridge: Medical Research Council.



# Obrigado!