# Introdução à modelagem estatística

Elias Teixeira Krainski
eliaskr@ufpr.br

Jul–2017, Lavras/MG
62ª RBras & 17º SEAGRO

# Exemplos de motivação

# Problemas e parâmetros de interesse

- concentração de GAG na urina de crianças em função da idade
- ocorrência de chuva em cada dia do ano
- mortalidade infantil

Parâmetros de interesse:

- concentração de GAG para uma dada idade
- probabilidade de chuva em cada dia
- taxa de mortalidade infantil - TMI

# Concentração de GAG na urina

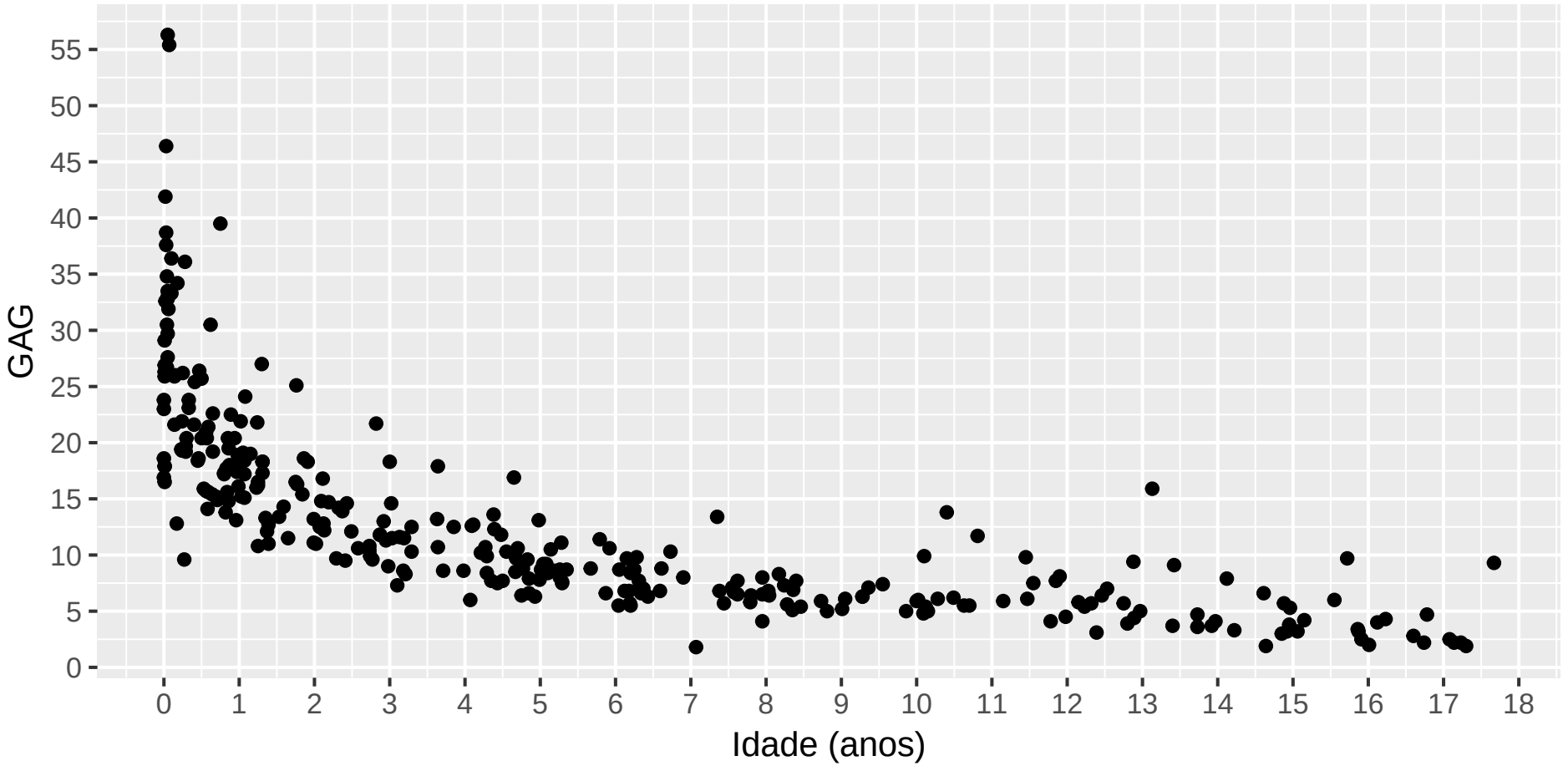



**Figura 1:** Concentração de GAG na urina de uma amostra de crianças sadias com as idades de interesse

# Dias chuvosos em Tokyo

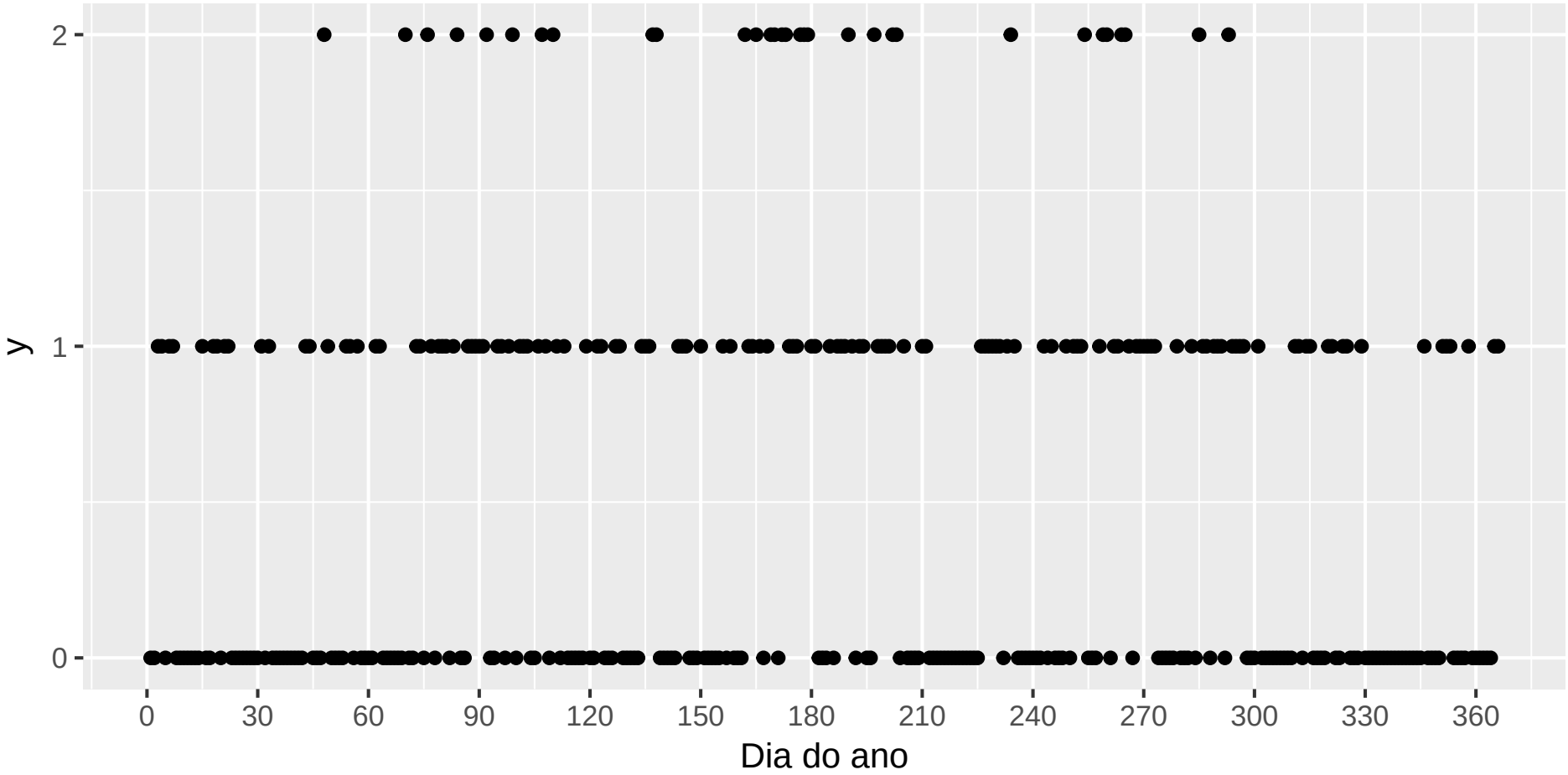
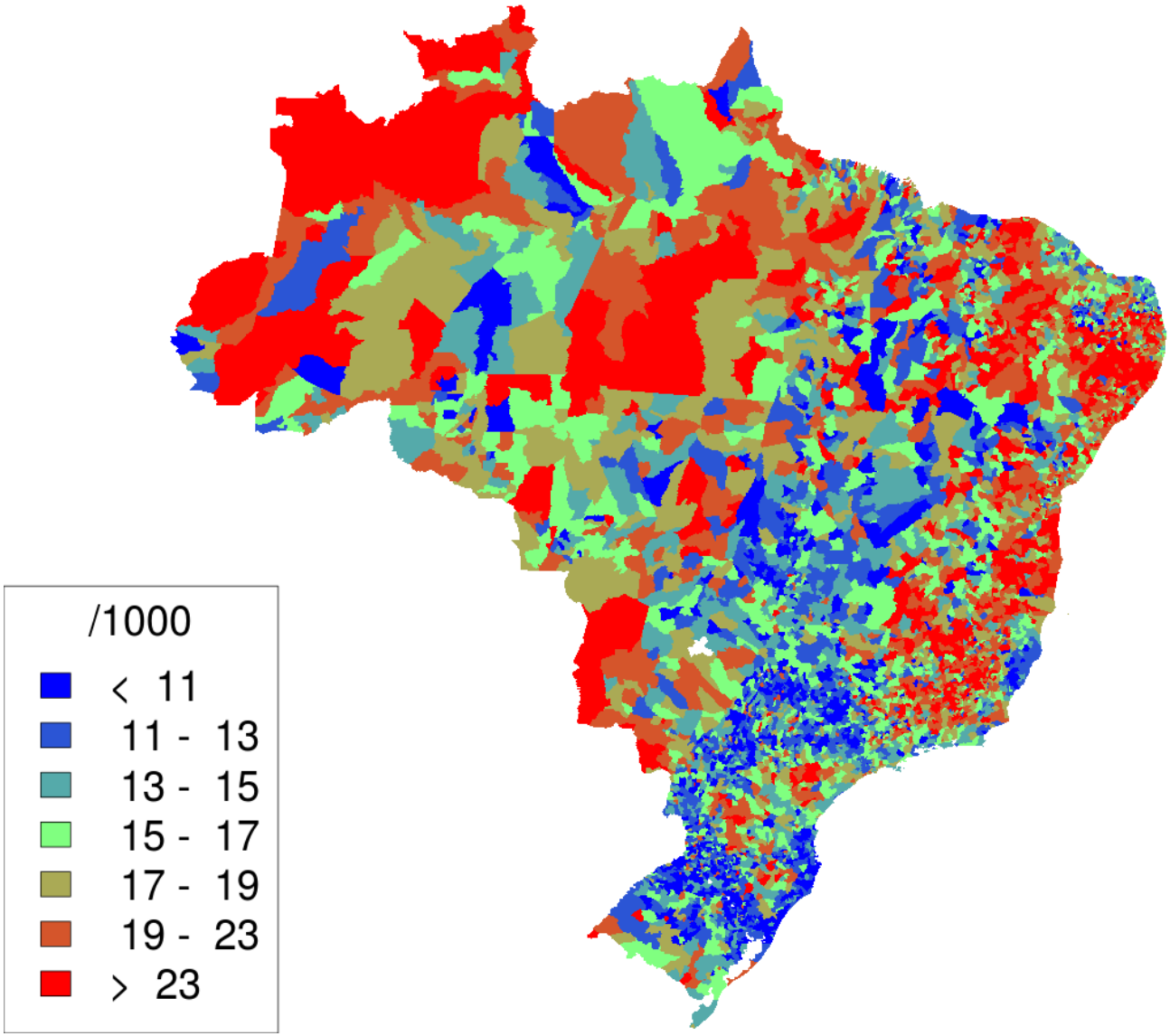



**Figura 2:** Choveu ou não em cada dia do ano durante dois anos.

''

# Taxa de mortalidade infantil por municípios



# Variabilidade amostral

# Dados observados

- pode ou não chover em primeiro de Janeiro de dois anos diferentes mesmo que a probabidade de chuva em primeiro de Janeiro não tenha alterado
- mesmo sem alteração nas condições de saúde de um município em dois anos consecutivos, implicando TMI igual nesses anos, a proporção (observada) de óbitos em cada ano pode ser diferente
- há crianças de cinco anos com GAG menor ou maior que 10, embora todas essas crianças sejam sadias

Variabilidade nos dados mesmo o parâmetro não se alterando

# Modelagem da variabilidade amostral

- assumir distribuição probabilística: **P( Dados | Parâmetros )**

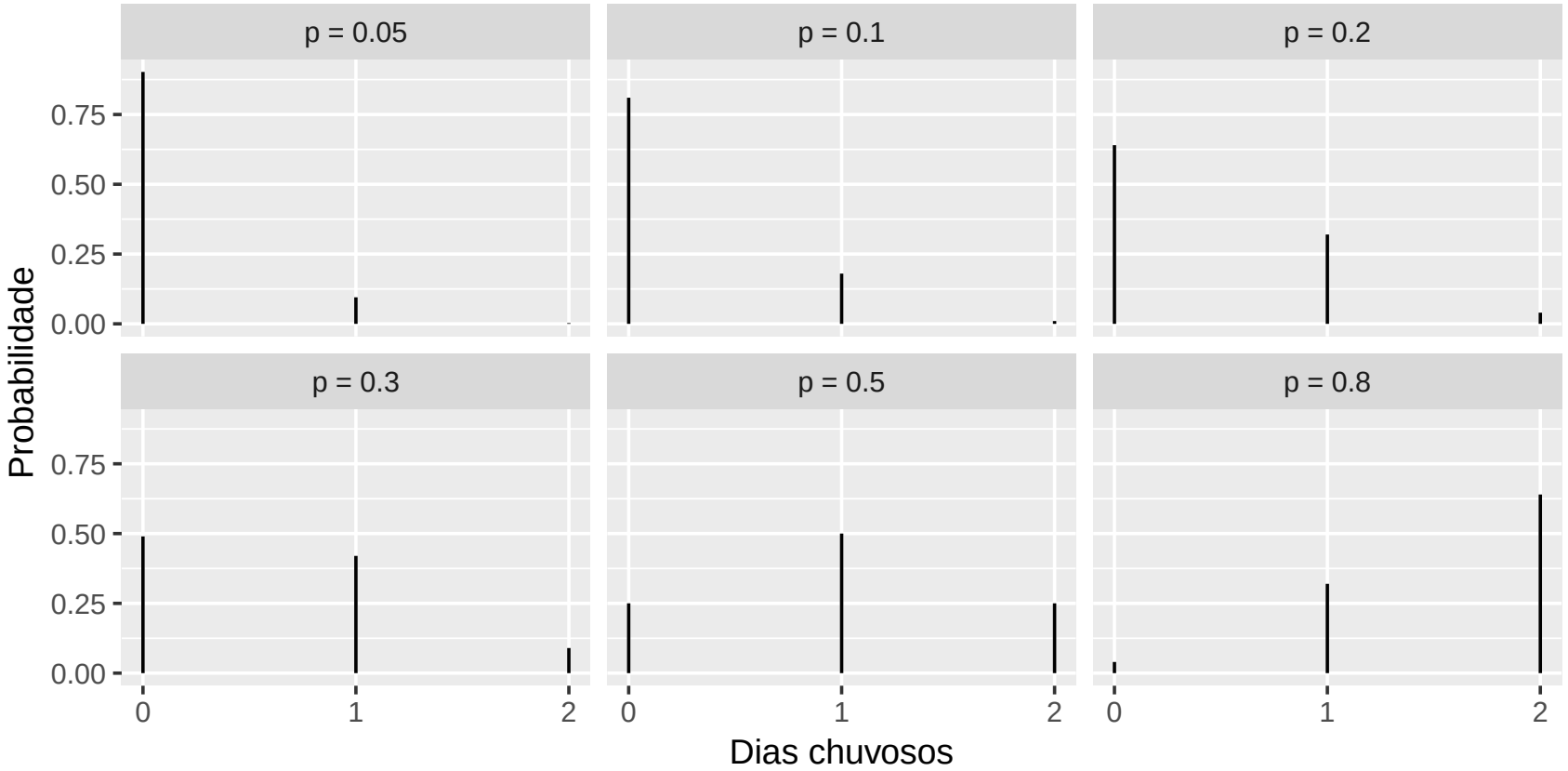
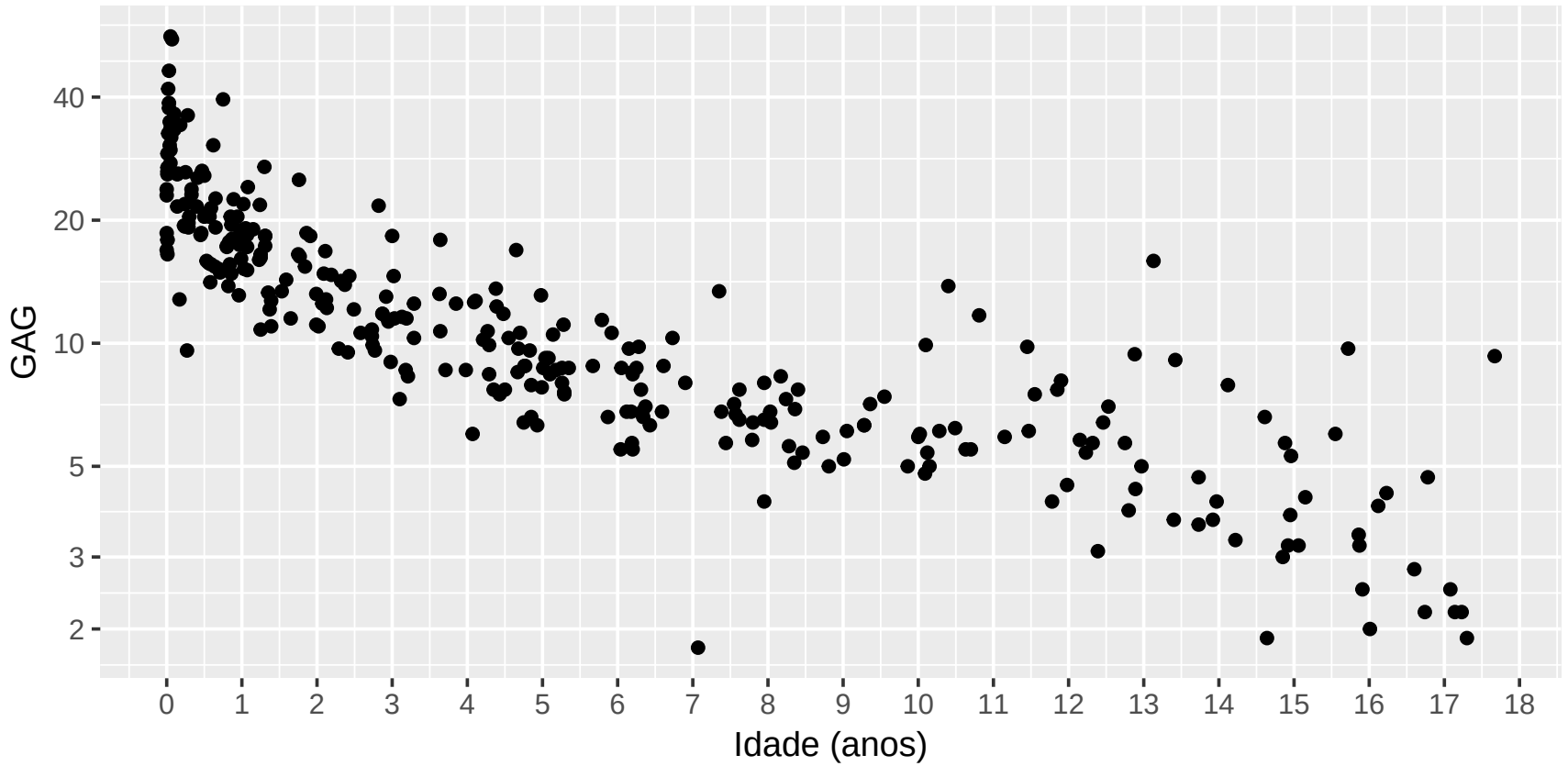



**Figura 3:** Binomial(2, p): Dias chuvosos em n=2 anos. Ex.: P(y=0|p=0.3)= 49%, P(y=1|p=0.3)=42% e P(y=2|p=0.3)=9%.

# Visualizando GAG na escala logaritma




- relação mais próxima da linear nessa escala que na original
- permite inferir GAG aos cinco anos mesmo sem crianças com cinco anos na amostra

# Uso de variáveis explicativas: Modelagem da média

- Podemos ter **P( Dados | Variáveis explicativas, Parâmetros)**
- valor esperado, ou média, para cada idade é o parâmetro de interesse
- sua relação com idade é importante e parametrizada
- logaritmo da média como função linear de idade

$$\log(\mu) = \beta_0 + \beta_1 \text{Idade}$$

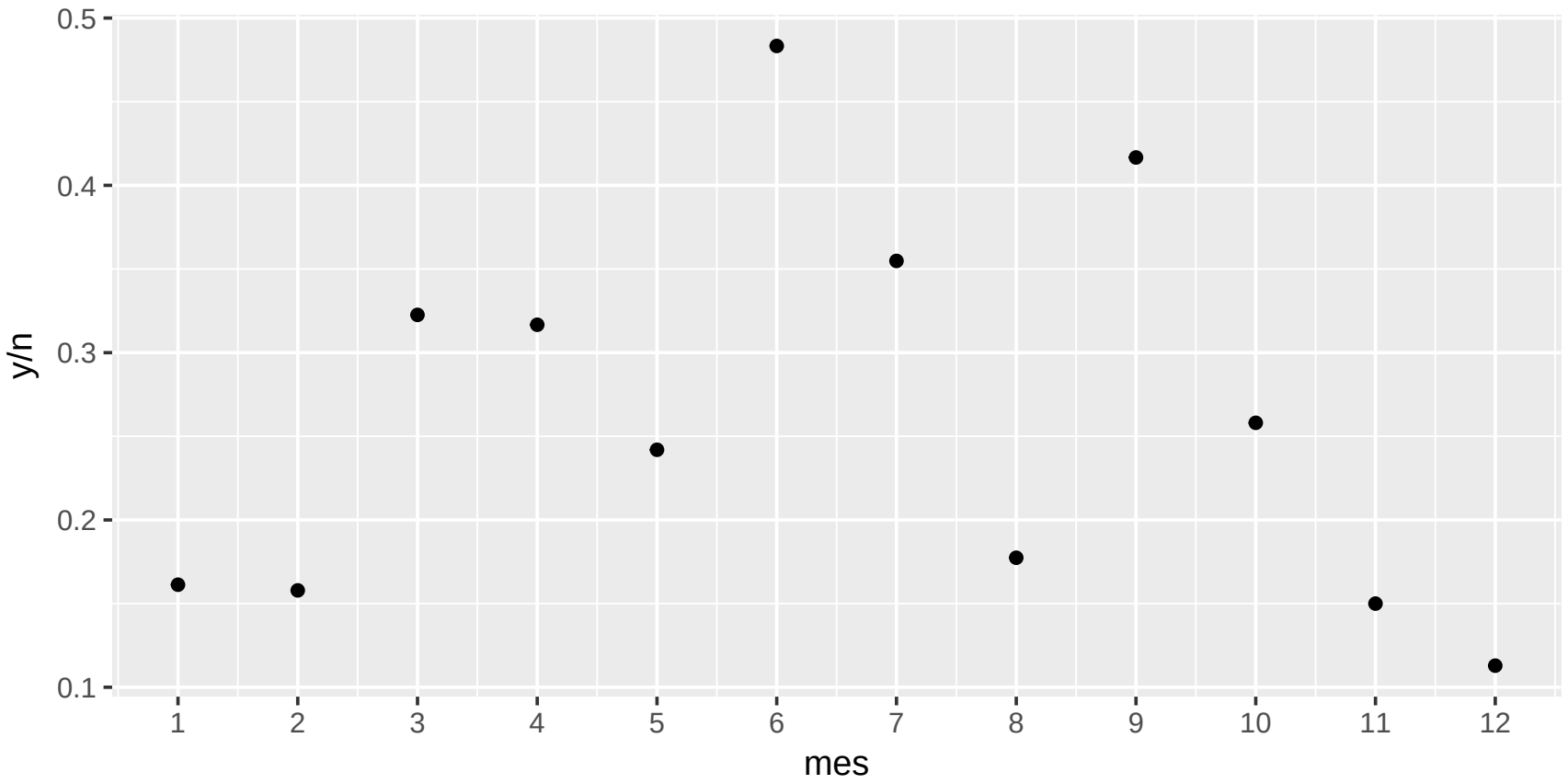
  onde $\beta_0$ e $\beta_1$ são parâmetros de interesse
- $e^{\beta_0}$: concentração esperada de GAG para Idade zero
- $\beta_1$: velocidade de decaimento

# Ocorrência de chuva em Tokyo

# Visualizar suavizações

- ferramenta exploratória
- há elegantes de suavização, exemplo: *Locally Weighted Scatterplot Smoothing* - LOWESS
- consideraremos opções mais simples a seguir

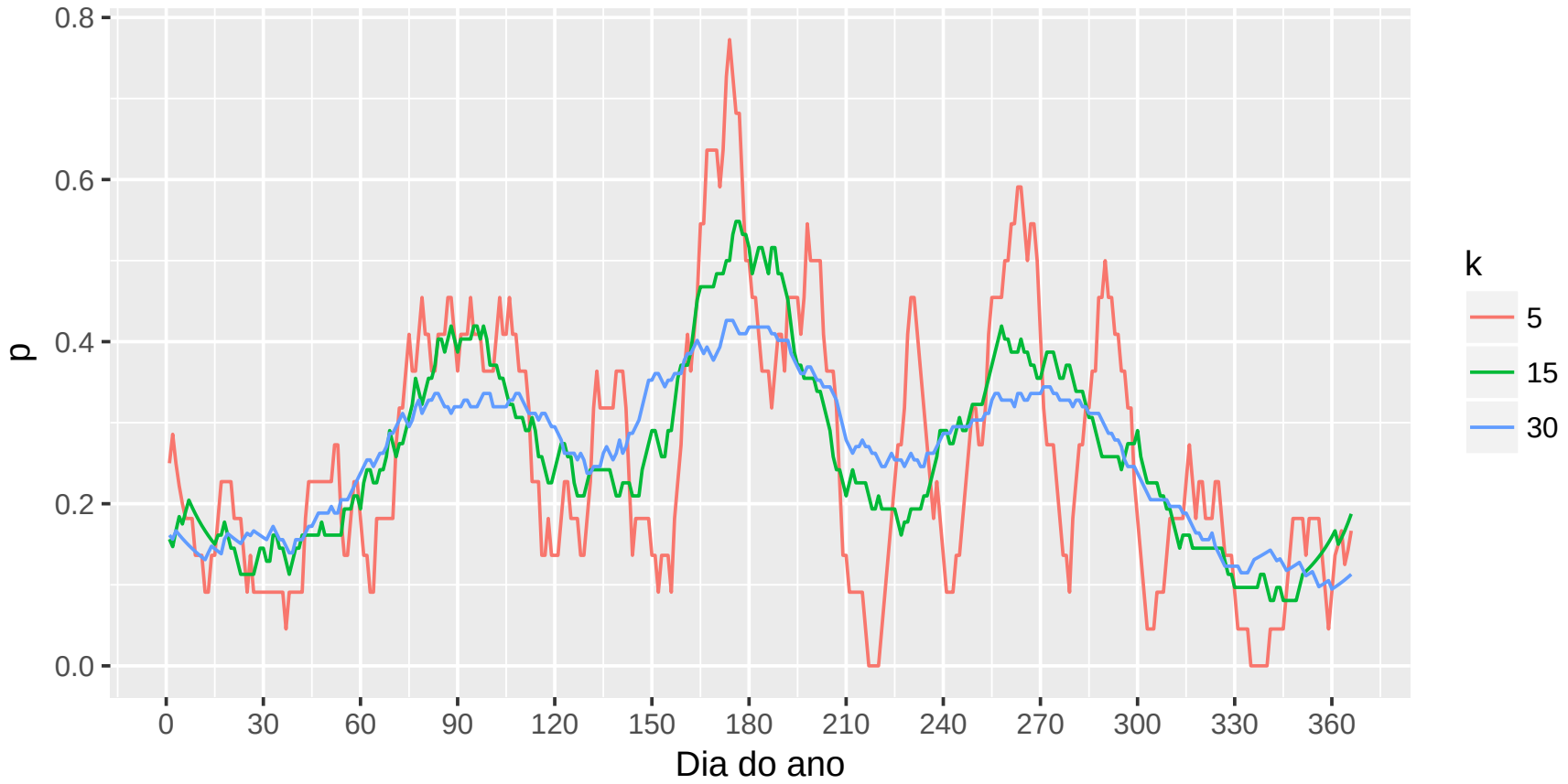
# Proporção de dias chuvosos por mês




- proporção de dias chuvosos não muda tão abruptamente
- porém, muito da variabilidade foi suprimida.

# Janela deslizante

- cada dia, a média dos dados nos dias mais próximos
- que estejam a uma distância menor que $k$ dias
- janelas deslizantes de amplitude $2 * k$ ao longo do ano



# Opções iniciais de modelagem: GLM

- probabilidade de chuva em função do tempo
- forma conveniente? $p$ ou numa escala transformada?
- modelos lineares generalizados, *Generalized Linear Models* - GLM
- função linear para o *logito* da probabilidade

$$p_i = \frac{1}{1 + \mathrm{e}^{-\eta_i}}$$

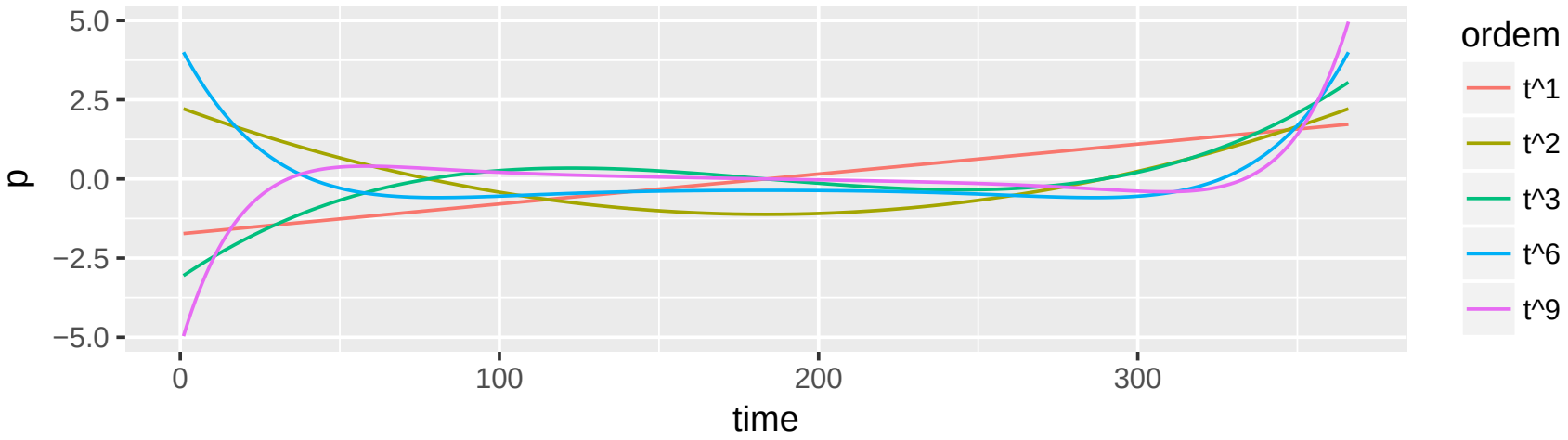
- qual função linear de tempo, $t$?
- **Savage**: "Devemos construir modelos tão grandes quanto elefantes"
- **von Neumann**: "Com quatro parâmetros eu posso estimar um elefante, e com cinco eu posso fazê-lo mexer sua tromba"

# Polinômios em $t$

- poliômio de ordem $m$

$$\eta_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + ... + \beta_m t_i^m$$

estimar os parâmetros $\beta_j$, $j = 1, ..., m$




- polinômios são flexíveis, mas carecem de interpretabilidade
- computacionalmente estável considerar polinômios ortogonais

# Resultado considerando polinômios

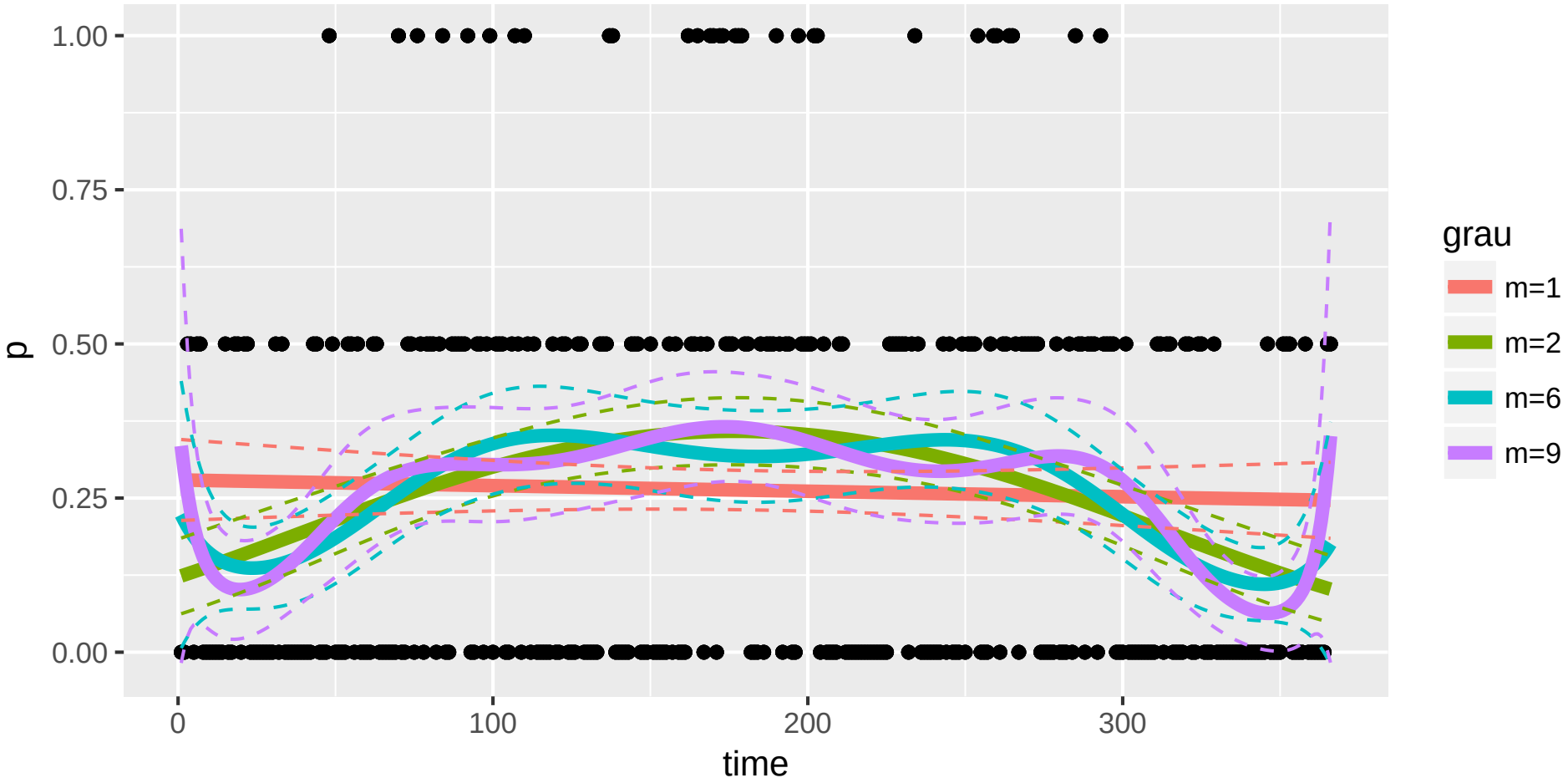



**Figura 4:** Curvas de predição (+ incerteza), para diferentes graus.

# Funções bases

- representar/subdividir o espaço da variável
- suporte compacto
  - cada função base representa uma parte
  - valores não nulos em parte da variável
  - coeficientes de regressão: ativação naquela parte

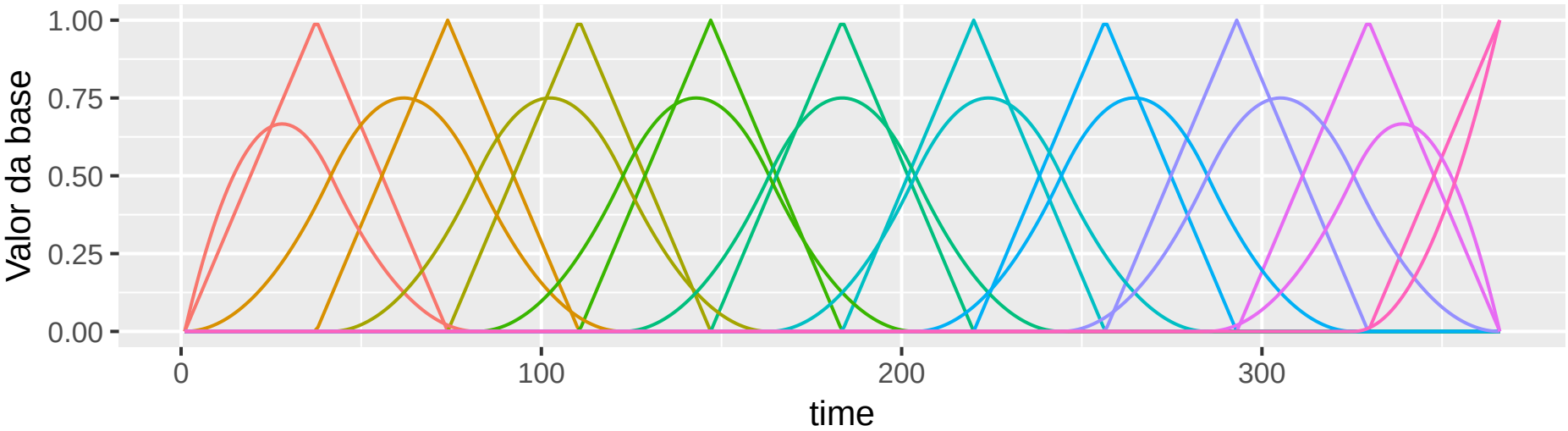



**Figura 5:** B-splines de primeiro e segundo grau, com 8 graus de liberdade.

# Usando B-splines de grau 2

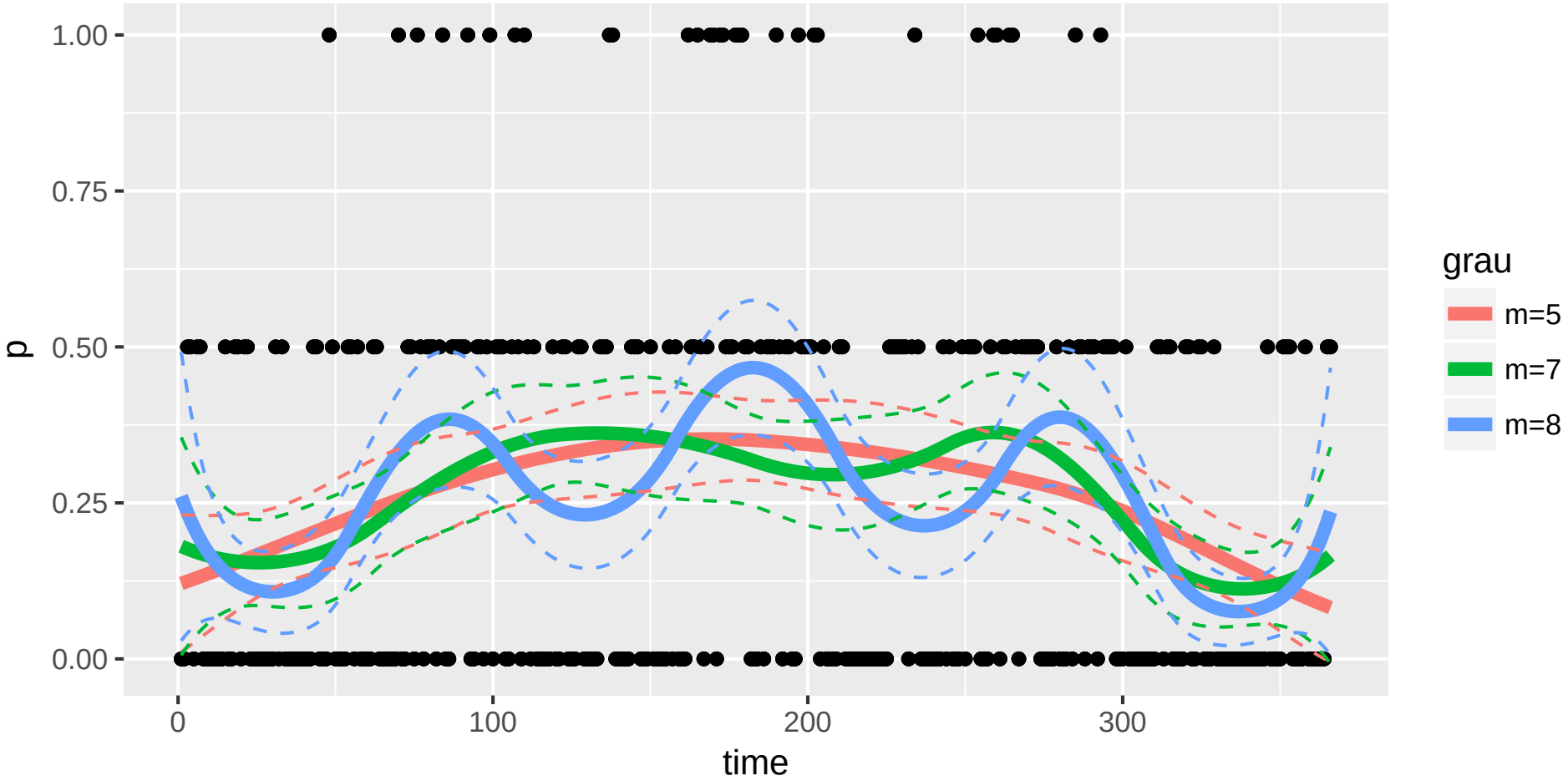
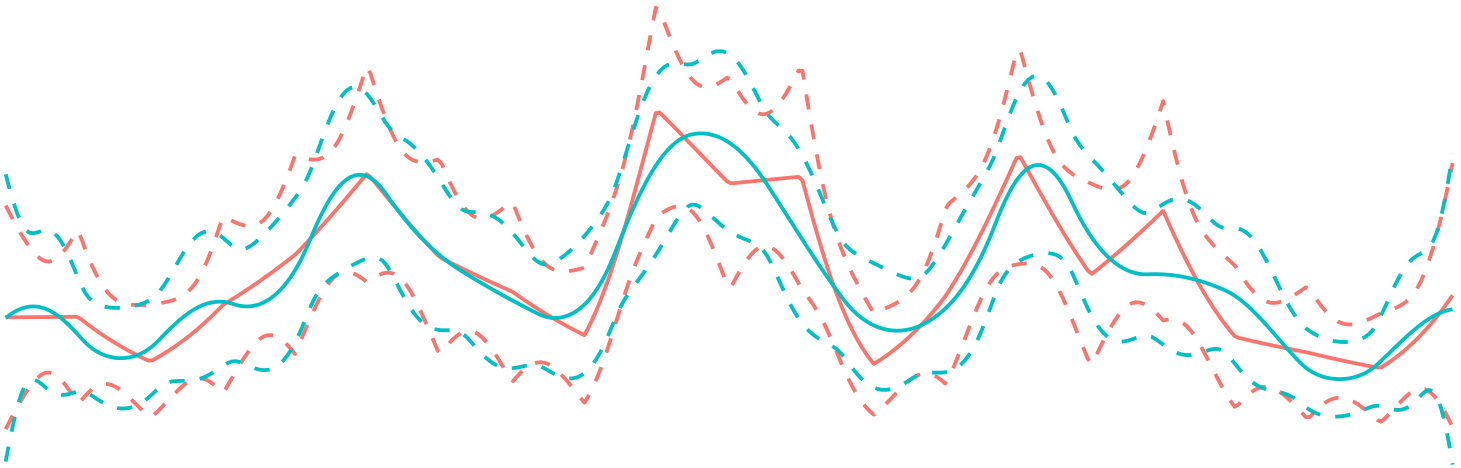


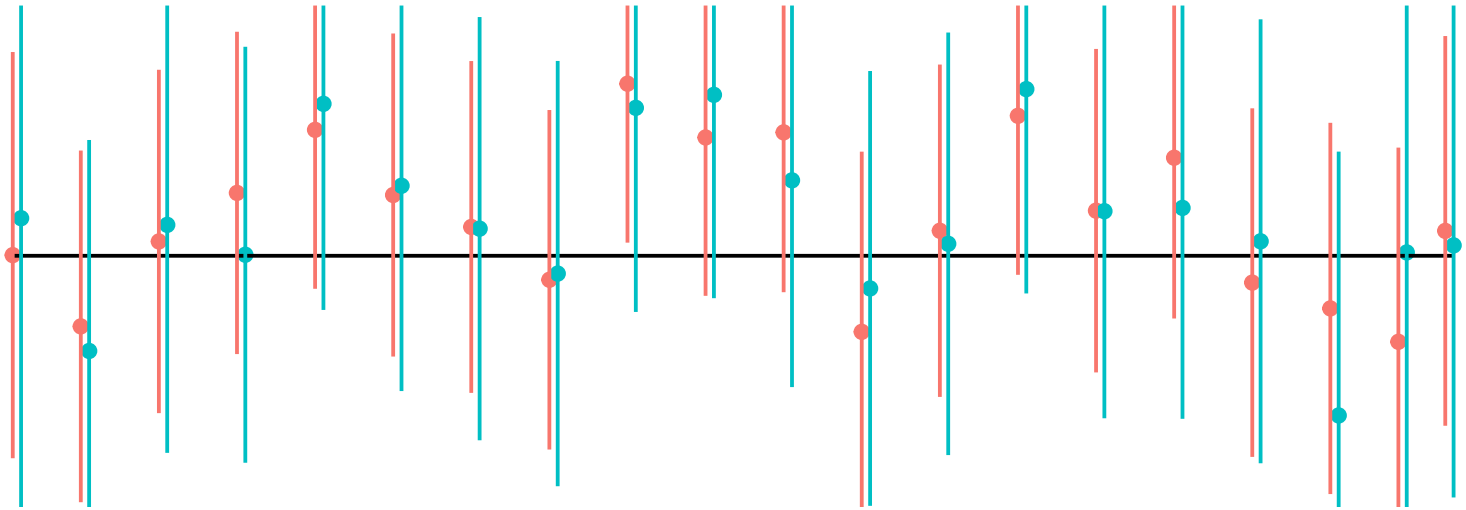

**Figura 6:** Curvas de predição (e bandas de incerteza), para diferentes graus de liberdade.

# Bases e coeficientes

# Exemplo: 20 *B-splines* de ordens 1 e 2




- O que ocorre com os coeficientes:



# Tempo discretizado (15 dias) como fator