# Modelagem Hierárquica Bayesiana para Análise do ECQ

---

*(Uma abordagem moderna para otimizar investimentos em rede com dados limitados)*

## Contexto do Problema

Na nossa rede, há vários sites em uma cidade, onde cada site possui um número variável de testes ECQ. Alguns sites têm poucos testes (ex: 5 a 20), enquanto outros têm centenas. A pergunta crítica é:

**Como identificar quais sites têm desempenho verdadeiramente abaixo do esperado e priorizar melhorias, sem ser enganado pela aleatoriedade de amostras pequenas?**

## Abordagens Tradicionais (e Seus Problemas)

### 1. Análise "Site por Site" (Sem Agrupamento)

**O que é:** Calcular a taxa de sucesso bruta para cada site

**Taxa Bruta = ECQ OK / TESTES_ECQ**

**Problemas:**

- Sites com poucos testes (ex: 2/5 sucessos = 40%) podem ter estimativas extremas e enganosas.
- Não diferencia variação real entre sites de ruído estatístico.

**Resultado:** Priorizações equivocadas (ex: investir em um site que parece ruim por acaso).

### 2. Agrupamento por cidade (Ignorar Diferenças)

**O que é:** Calcular uma taxa de sucesso única para todos os sites da cidade (ex: média geral = 75%).

**Problemas:**

- Ignora variações reais entre sites

**Resultado:** Falha em identificar problemas localizados, prejudicando a otimização da rede.

## Solução Proposta: Modelagem Hierárquica Bayesiana

### Como Funciona

A técnica combina dados individuais de cada site com informações da população total de sites, usando um modelo estatístico de dois níveis:

- **Nível 1 (Site):** Estima a taxa de sucesso real theta_j para cada site.
- **Nível 2 (Cidade):** Aprende a distribuição geral de sucesso média m e a variabilidade entre sites phi.
- **Nível 3 (ANF).**

## Benefícios-Chave

✅ **Estimativas Estáveis:**

- Sites com poucos testes têm suas estimativas "puxadas" para a média populacional, reduzindo falsos alarmes.
- Sites com muitos testes mantêm estimativas próximas dos dados observados.

✅ **Quantificação da Incerteza:**

- Fornece intervalos de credibilidade (ex: "95% de chance da taxa real estar entre 65% e 82%").
- Evita decisões baseadas em dados insuficientes.

✅ **Identificação de Padrões:**

- Detecta se a variação entre sites é alta (ex: alguns sites são sistematicamente piores) ou baixa (ex: desempenho homogêneo). ## Exemplo Prático: Comparação de Abordagens

| Site | Testes | Sucessos | Taxa Bruta | Modelo Hierárquico (Intervalo 95%) |
|---|---|---|---|---|
| A | 5 | 2 | 40% | 28% – 65% |
| B | 100 | 85 | 85% | 77% – 91% |

- **Site A (5 testes):** A taxa bruta (40%) sugere problema, mas o modelo mostra que a taxa real pode ser até 65% (intervalo amplo). Investir aqui pode ser um gasto desnecessário.
- **Site B (100 testes):** Alta confiança na taxa de 85%. Priorizar outros sites. ## Impacto Gerencial

## Redução de Custos:

- Evita investimentos em sites que parecem ruins devido a amostras pequenas. ### Otimização de Recursos:

- Foca em sites com baixo desempenho real (ex: intervalo de credibilidade abaixo de 60%).

## Tomada de Decisão Informada:

- Relatórios claros com níveis de confiança, facilitando a justificativa para stakeholders

# Recomendação

Adote a modelagem hierárquica Bayesiana para:

- Transformar dados esparsos em insights confiáveis e acionáveis.
- Equilibrar a flexibilidade de análises individuais com a robustez do agrupamento inteligente.

Esta abordagem é amplamente utilizada por operadoras líderes globais e está alinhada com as melhores práticas de data science aplicado a telecomunicações.

# Exemplo de Modelo com 3 Níveis (Site → Cidade → Estado)

Suponha que você tenha:

- **Nível 1 (Site):** ECQ tests de cada site.
- **Nível 2 (Cidade):** Sites agrupados por cidades.
- **Nível 3 (Estado):** Cidades agrupadas por estados.

## Estrutura do Modelo

**Nível 1 (Site):**

Para cada site $k$ na cidade $j$:

$$n\_success_{jk} \sim Binomial(n\_tests_{jk}, \theta_{jk})$$

Onde:

- $n\_success_{jk}$: Número de testes bem-sucedidos no site $k$ da cidade $j$.
- $n\_tests_{jk}$: Número total de testes no site $k$ da cidade $j$.
- $\theta_{jk}$: Taxa de sucesso do site $k$ na cidade $j$.

**Nível 2 (Cidade):**

Os sites de uma cidade compartilham uma média regional:

$$\theta_{jk} \sim Beta(\mu_j \phi_j, (1-\mu_j)\phi_j)$$

Onde:

- $\mu_j$: Taxa média de sucesso da cidade $j$.
- $\phi_j$: Precisão (variabilidade entre sites na cidade $j$).

**Nível 3 (Estado):**

As cidades de um estado compartilham uma média global:

$$\mu_j \sim Beta(\alpha_{estado}, \beta_{estado})$$

$$\phi_j \sim Half-Normal(0, \sigma_{estado})$$

Onde:

- $\alpha_{estado}, \beta_{estado}$: Hiperparâmetros do estado.
- $\sigma_{estado}$: Variabilidade entre cidades no estado.

## Benefícios da Abordagem Multinível

- **Identificação de Padrões Geográficos:**
  - Descubra se problemas são **localizados** (ex: uma cidade específica) ou **sistêmicos** (ex: todo o estado).
  - Exemplo: Se todas as cidades do Estado X têm baixo desempenho, investigue causas macro (ex: infraestrutura estadual).
- **Alocação de Recursos Estratégica:**
  - Compare estados/cidades para priorizar investimentos onde o impacto será maior.
  - Exemplo: Se o Estado A tem cidades com desempenho consistentemente baixo, priorize upgrades de infraestrutura regional.
- **Estabilidade em Dados Escassos:**
  - Cidades com poucos sites (ou estados com poucas cidades) "herdam" informação do nível superior.
  - Exemplo: Uma cidade com 2 sites (poucos dados) terá estimativas de $\mu_j$ regulareshadas para a média do estado.

## Dados de ECQ da TNE de Fevereiro / 2025

---

Este é um modelo hierárquico bayesiano sofisticado de três níveis:

- Nível 3 (ANF): Define distribuições para cada ANF
- Nível 2 (Município): Cada município tem uma distribuição cujos parâmetros dependem da ANF correspondente
- Nível 1 (Site): Cada site/observação tem uma distribuição que depende do município correspondente

O modelo é estruturado para capturar a variabilidade em diferentes níveis e permitir estimativas de parâmetros em todos os níveis da hierarquia. É particularmente útil para identificar a variação na qualidade ou conformidade entre diferentes regiões administrativas e municípios. O uso de distribuições Beta em cada nível é apropriado para modelar proporções (taxas de sucesso), e os parâmetros phi permitem controlar a concentração dessas distribuições, possibilitando um modelo flexível.

In [ ]:
```
import pandas as pd
import pymc as pm
import os
os.environ["MKL_NUM_THREADS"] = "8"
os.environ["OMP_NUM_THREADS"] = "8"
os.environ["NUMBA_NUM_THREADS"] = "8"

# Carregar dados
data = pd.read_excel('ECQ_FEV25.xlsx')

# Agrupar por ANF e Município (DEFINA AQUI)
grouped = data.groupby(['ANF', 'MUNICIPIO'])
```

```
for name, group in grouped:
    print(f"Group name: {name}")
    print(group)
```

# Como o Modelo Hierárquico Funciona

O modelo utiliza a estrutura hierárquica dos dados brutos (não agregados) para estimar:

**Nível Site:**

Taxa de sucesso de cada site:

$$\theta_{\text{site}} \sim \text{Beta}(\mu_{\text{cidade}} \cdot \phi, (1 - \mu_{\text{cidade}}) \cdot \phi)$$

**Nível Cidade:**

Taxa média da cidade:

$$\mu_{\text{cidade}} \sim \text{Beta}(\alpha_{\text{estado}}, \beta_{\text{estado}})$$

**Nível Estado:**

Hiperparâmetros do estado:

$$\alpha_{\text{estado}}, \beta_{\text{estado}} \sim \text{Prior}$$

## Estrutura do Modelo Hierárquico ECQ (ANF, Município, ENDERECO_ID)

---

### Contexto

Você deseja estimar a taxa de sucesso de testes ECQ em **sites 4G**, considerando a estrutura hierárquica:

- **ANF** (nível mais alto: região geográfica).
- **Município** (subdivisão da ANF).
- **ENDERECO_ID** (site específico dentro de um município).

Muitos sites têm poucos testes, e a modelagem hierárquica permite:

- Regularizar estimativas de sites com poucos dados.
- Identificar padrões de desempenho em nível de município e ANF.

---

### Modelo em Três Níveis

1. **Nível 1 (Sites dentro de Municípios)**
   Para cada site (k) no município ( j):

$$\text{TESTES\_ECQ\_OK}_{jk} \sim \text{Binomial}(\text{TESTES\_ECQ}_{jk}, \theta_{jk})$$

- **Distribuição Binomial**: Modela o número de sucessos ( $\text{TESTES\_ECQ\_OK}$) em $\text{TESTES\_ECQ}$ tentativas.

- **Parâmetro**:
  $\theta_{jk}$: Taxa de sucesso verdadeira do site $k$ no município $j$.

1. **Nível 2 (Municípios dentro de ANFs)**
   As taxas de sucesso dos sites em um município são agrupadas em torno de uma média municipal:

$$\theta_{jk} \sim \text{Beta}(\mu_j \phi_j, (1 - \mu_j)\phi_j)$$

   - **Distribuição Beta**: Modela probabilidades (0–1) e é conjugada com a Binomial.
   - **Parâmetros**:
     - $\mu_j$: Taxa média de sucesso do município $j$.
     - $\phi_j$: Precisão (controla a variabilidade entre sites no município).
2. **Nível 3 (ANFs)**
   As médias dos municípios são agrupadas em torno de uma média regional (ANF):

$$\mu_j \sim \text{Beta}(\mu_{\text{ANF}} \phi_{\text{ANF}}, (1 - \mu_{\text{ANF}})\phi_{\text{ANF}})$$

   - **Hiperparâmetros**:
     - $\mu_{\text{ANF}}$: Taxa média de sucesso da ANF.
     - $\phi_{\text{ANF}}$: Precisão (controla variabilidade entre municípios na ANF).

---

### Priors para Hiperparâmetros

[

$$\begin{aligned} \mu_{\text{ANF}} &\sim \text{Beta}(2, 2) && (\text{Prior fracamente informativa, centrada em } 0.5) \\ \phi_{\text{ANF}} &\sim \text{HalfNormal}(0, 10) && (\text{Prior que regulariza a variabilidade}) \end{aligned}$$

]

---

## Por Que Essas Distribuições?

| Componente | Distribuição | Motivo |
|---|---|---|
| **Dados (TESTES_ECQ_OK)** | Binomial | Contagens de sucessos em ensaios independentes (sucesso/falha). |
| **Taxa de sucesso ((\theta))** | Beta | Natural para probabilidades e conjugada com Binomial. |
| **Média municipal ((\mu_j))** | Beta | Mantém a coerência com a escala de probabilidade (0–1). |
| **Precisão ((\phi))** | Half-Normal | Garante valores positivos e evita overfitting. |
| **Média da ANF ((\mu_{\text{ANF}}))** | Beta(2,2) | Prior flexível, equivalente a 4 "observações fictícias" (2 sucessos + 2 falhas). |

---

## Fluxo do Modelo

1. **ANF → Município**:
   - Cada ANF tem uma distribuição de desempenho ($\mu_{\text{ANF}}, \phi_{\text{ANF}}$).

- Municípios "herdam" essa distribuição, com suas próprias médias ($\mu_j$).

2. **Município → Site**:
   - Sites dentro de um município compartilham a média $\mu_j$, mas podem variar (controlado por $\phi_j$).
3. **Resultado**:
   - Sites com poucos testes são regularizados em direção à média do município/ANF.
   - Sites com muitos testes mantêm estimativas próximas dos dados observados.

---

## Equações Consolidados

[

$$
\begin{aligned}
\text{TESTES\_ECQ\_OK}_{jk} &\sim \text{Binomial}(n_{jk}, \theta_{jk}) && \text{(Likelihood)} \\
\theta_{jk} &\sim \text{Beta}(\mu_j\phi_j, (1-\mu_j)\phi_j) && \text{(Prior dos sites)} \\
\mu_j &\sim \text{Beta}(\mu_{\text{ANF}}\phi_{\text{ANF}}, (1-\mu_{\text{ANF}})\phi_{\text{ANF}}) && \text{(Prior dos municíp} \\
\mu_{\text{ANF}} &\sim \text{Beta}(2, 2) && \text{(Prior da ANF)} \\
\phi_{\text{ANF}} &\sim \text{HalfNormal}(0, 10) && \text{(Prior da precisão)}
\end{aligned}
$$

]

---

## Visualização da Hierarquia

In [ ]:
```python
with pm.Model() as modelo_flex_phi_estimado:
    # Hiperpriors para ANFs (Nível 3)
    anfs = data['ANF'].unique()
    n_anfs = len(anfs)

    mu_anf = pm.Beta("mu_anf", alpha=2, beta=2, shape=n_anfs)
    sigma_anf = pm.HalfNormal("sigma_anf", sigma=0.1, shape=n_anfs)

    # Phi estimado para municípios e sites
    phi_municipio = pm.HalfNormal("phi_municipio", sigma=10)  # Nível 2
    phi_site = pm.HalfNormal("phi_site", sigma=10)            # Nível 1

    # Mapeamento ANF -> índice
    anf_id_map = {anf: idx for idx, anf in enumerate(anfs)}

    # Loop por cada grupo (ANF + Município)
    theta_sites_list = []
    for (anf, municipio), group in grouped:
        # Dados do município
        n_tests = group['TESTES_ECQ'].values
        n_success = group['TESTES_ECQ_OK'].values
        idx_anf = anf_id_map[anf]

        # Nível 2: Município (mu_municipio ~ ANF)
        mu_municipio = pm.Beta(
            f"mu_municipio_{anf}_{municipio}",
            alpha=mu_anf[idx_anf] * phi_municipio,
            beta=(1 - mu_anf[idx_anf]) * phi_municipio
        )
```

```
        # Nível 1: Sites (theta_site ~ município)
        theta_site = pm.Beta(
            f"theta_site_{anf}_{municipio}",
            alpha=mu_municipio * phi_site,
            beta=(1 - mu_municipio) * phi_site,
            shape=len(n_tests)
        )

        # Likelihood
        pm.Binomial(
            f"obs_{anf}_{municipio}",
            n=n_tests,
            p=theta_site,
            observed=n_success)

    # Amostragem
    trace = pm.sample(2000, tune=1000, chains=2, target_accept=0.9)
```

In [ ]:
```
import arviz as az

az.summary(trace, var_names=["mu_anf", "mu_municipio", "theta_site"])
# Comparar ANFs
az.plot_forest(trace, var_names="mu_anf", combined=True)

# Detalhar um município específico
az.plot_forest(trace, var_names="mu_municipio_82_ANADIA", combined=True)

# Ver sites de um município
az.plot_forest(trace, filter_vars="like", var_names="theta_site_82_ANADIA", combine
```

# Análise Gerencial usando Modelagem Hierárquica Bayesiana

A modelagem hierárquica Bayesiana é uma ferramenta poderosa para transformar dados operacionais em **insights estratégicos**. Vou detalhar como você pode utilizar esse modelo em análises gerenciais, com exemplos práticos e perguntas que o modelo pode responder.

---

## 1. Análise de Desempenho Agregado

### Pergunta: Qual ANF tem a pior taxa de sucesso agregada?

- **Como responder**:
  Compare as distribuições posteriores de `mu_anf` (média de sucesso por ANF).
  - Use `az.summary(trace, var_names=["mu_anf"])` para obter médias e intervalos de credibilidade.
  - Visualize com `az.plot_forest(trace, var_names=["mu_anf"])`.
- **Exemplo de Insight**:

> *"A ANF 82 tem a menor taxa média de sucesso (65%), com 95% de credibilidade entre 60% e 70%. Recomenda-se investigar causas estruturais nessa região."*

---

## 2. Identificação de Outliers

### Pergunta: Há municípios na ANF X que são outliers (para melhor ou pior)?

- **Como responder**:
  Analise as posteriores de `mu_municipio` para municípios dentro da ANF X.
  - Use `az.summary(trace, var_names=["mu_municipio"])` para identificar municípios com médias extremas.
  - Visualize com `az.plot_forest(trace, var_names=["mu_municipio"])`.
- **Exemplo de Insight**:

> *"Na ANF 82, o município de Arapiraca tem uma taxa de sucesso de 85% (acima da média da ANF), enquanto Água Branca tem apenas 50%. Recomenda-se investigar boas práticas em Arapiraca e problemas em Água Branca."*

---

## 3. Priorização de Intervenções Locais

### Pergunta: Qual site na Cidade Y merece intervenção imediata?

- **Como responder**:
  Verifique as posteriores de `theta_site` para sites na Cidade Y.
  - Use `az.summary(trace, var_names=["theta_site"])` para identificar sites com baixo desempenho.
  - Visualize com `az.plot_forest(trace, var_names=["theta_site"])`.
- **Exemplo de Insight**:

> *"Na Cidade Y, o site ALABN_0002 tem uma taxa de sucesso de 40% (intervalo de credibilidade: 35%–45%), significativamente abaixo da média da cidade (70%). Recomenda-se uma inspeção técnica urgente."*

---

## 4. Comparação de Desempenho entre Regiões

### Pergunta: Como o desempenho da ANF 82 se compara à ANF 83?

- **Como responder**:
  Compare as posteriores de `mu_anf` para as duas ANFs.
  - Use `az.plot_posterior(trace, var_names=["mu_anf"])` para visualizar as distribuições.
  - Calcule a probabilidade de uma ANF ser melhor que a outra:

```
prob = (trace.posterior["mu_anf"][0] > trace.posterior["mu_anf"]
[1]).mean()
print(f"Probabilidade de ANF 82 ser melhor que ANF 83:
{prob:.2%}")
```

- **Exemplo de Insight**:

> *"A ANF 82 tem 90% de probabilidade de ser pior que a ANF 83. Recomenda-se alocar recursos adicionais para melhorar a infraestrutura na ANF 82."*

---

## 5. Análise de Tendências Temporais

### Pergunta: O desempenho da ANF 82 melhorou ao longo do tempo?

- **Como responder**:
  - Divida os dados por período (ex: trimestres) e ajuste o modelo separadamente para cada período.
  - Compare as posteriores de `mu_anf` ao longo do tempo.
- **Exemplo de Insight**:

> *"A taxa de sucesso da ANF 82 aumentou de 60% no Q1 para 70% no Q4, indicando que as intervenções recentes estão surtindo efeito."*

---

## 6. Identificação de Fatores de Sucesso

### Pergunta: Quais fatores estão associados a um melhor desempenho?

- **Como responder**:
  - Adicione variáveis explicativas ao modelo (ex: tipo de antena, densidade populacional).
  - Verifique os coeficientes posteriores dessas variáveis.
- **Exemplo de Insight**:

> *"Sites com antenas modernas têm uma taxa de sucesso 15% maior que sites com antenas antigas. Recomenda-se priorizar upgrades de infraestrutura."*

---

## 7. Simulação de Cenários

### Pergunta: Qual seria o impacto de melhorar os 10% piores sites?

- **Como responder**:
  - Simule um cenário onde os 10% piores sites ( `theta_site` ) são ajustados para a média da ANF.
  - Recalcule a taxa de sucesso agregada.
- **Exemplo de Insight**:

> *"Melhorar os 10% piores sites aumentaria a taxa de sucesso da ANF 82 de 65% para 75%, impactando positivamente a experiência de 50.000 usuários."*

---

## 8. Alocação de Recursos

### Pergunta: Onde alocar recursos para maximizar o impacto?

- **Como responder**:
  - Combine as análises de desempenho agregado, outliers e fatores de sucesso.
  - Priorize regiões com baixo desempenho e alto potencial de melhoria.
- **Exemplo de Insight**:

> "*Recomenda-se alocar 70% do orçamento para a ANF 82 (baixo desempenho) e 30% para a ANF 83 (manutenção preventiva).*"

---

## 9. Relatórios Gerenciais

### Exemplo de Relatório:

1. **Desempenho Agregado**:
   - ANF 82: 65% (60%–70%)
   - ANF 83: 75% (70%–80%)
2. **Outliers**:
   - Melhor município: Arapiraca (85%)
   - Pior município: Água Branca (50%)
3. **Sites Críticos**:
   - ALABN_0002: 40% (35%–45%)
   - ALAIR_0001: 87% (85%–90%)
4. **Recomendações**:
   - Investir em upgrades de antenas na ANF 82.
   - Realizar inspeções técnicas em Água Branca.
   - Replicar boas práticas de Arapiraca em outras cidades.

---

## Ferramentas para Visualização e Apresentação

- **ArviZ**: Para gráficos de posteriores, intervalos de credibilidade e comparações.
- **Power BI**: Para dashboards interativos com os resultados do modelo.
- **Relatórios PDF**: Com gráficos e insights gerados automaticamente.