Introdução à Inferência Estatística

Prof. Caio Azevedo

Introdução

- Estatística: área do conhecimento/Ciência que trata de metodologias (matemáticas) apropriadas para se coletar, organizar e analisar dados.
- A Estatística é uma ferramenta muito importante na resolução de problemas levantados nas mais diversas áreas do conhecimento: Biologia, Psicometria, Educação, Medicina, Física, Computação entre outras.
- É importante que o Estatístico participe de todas as etapas de um estudo (pesquisa/consultoria).

Etapas para a resolução de um problema

1. Determinação do problema/objeto de estudo.
2. Determinação dos objetivos específicos.
3. Determinação do tamanho da amostra.
4. Execução do levantamento dos dados: amostragem, entrevistas, experimento, coleta de dados etc.
5. Análise Descritiva.
6. **Análise Inferencial** .
7. Conclusões e elaboração dos relatórios/artigos/trabalhos pertinentes.

Pode-se retornar a pontos anteriores ou mesmo avançar (pulando alguns), consoante a necessidade.

Notações

- Variável aleatória: quantidade desconhecida e observável, representada por uma letra (latina) maiúscula: X, Y, Z. Respectivo valor observado, representado por uma letra minúscula: x, y, z.
- Parâmetro: objeto de interesse, quantidade desconhecida e não observável, representado por uma letra grega minúscula: θ, γ, β.
- Vetor (negrito). Vetor aleatório $\boldsymbol{X} = (X_1, ..., X_n)'$; vetor paramétrico $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ (vetores coluna).
- Tamanho da amostra: n. Tamanho da população: N (em geral, trabalharemos com populações de tamanho infinito ou inifinitamente grande).

Notação

- Número de parâmetros: p (eventualmente, p = 1).
- Função de probabilidade (caso discreto) ou função densidade (de probabilidade), ou simplesmente densidade (caso contínuo) (fdp): $f_X(.) \equiv f_X(.;\theta) \equiv p_X(.) \equiv p_X(.;\theta)$ (o subíndice pode ser, eventualmente suprimido).
- Função de distribuição acumulada (fda): $F_X(.) \equiv F_X(.;\theta)$ (o subíndice pode ser, eventualmente suprimido).
- Quando do cálculo das duas quantidades acima, em algum valor de interesse, digamos "x", o símbolo "." será substituído por este (x).

Problema probabilístico

- Considere um espaço probabilístico $(\Omega, \mathcal{A}, \mathcal{P})$, em que :
  - Ω : espaço amostral -conjunto que contempla todos os resultados possíveis de um experimento aleatório.
  - $\mathcal{A}$: $\sigma-$álgebra - conjunto que contem todos os subconjuntos de (interesse de) Ω.
  - $\mathcal{P}$: medida de probabilidade (ou probabilidade) - $\mathcal{P} : \mathcal{A} \rightarrow [0, 1]$.

Problema probabilístico

- Objetivo(s): deseja-se calcular quantidades de interese como: média, variância, probabilidades, funções de probabilidade/densidade, funções geradoras de momento (fgm) etc.
- Em geral $\mathcal{P}$ corresponde a alguma(s) fda(s) (função de distribuição acumulada), digamos $F_X(.)$, associada à alguma variável aleatória (va), digamos X.
- Tal fda é (suposta) conhecida tanto em relação à sua forma (normal, gama, Poisson, binomial etc) quanto em relação aos seus parâmetros (média, variância, parâmetro de forma, assimetria etc).
- Podemos ter mais de uma fda (va) envolvidas.

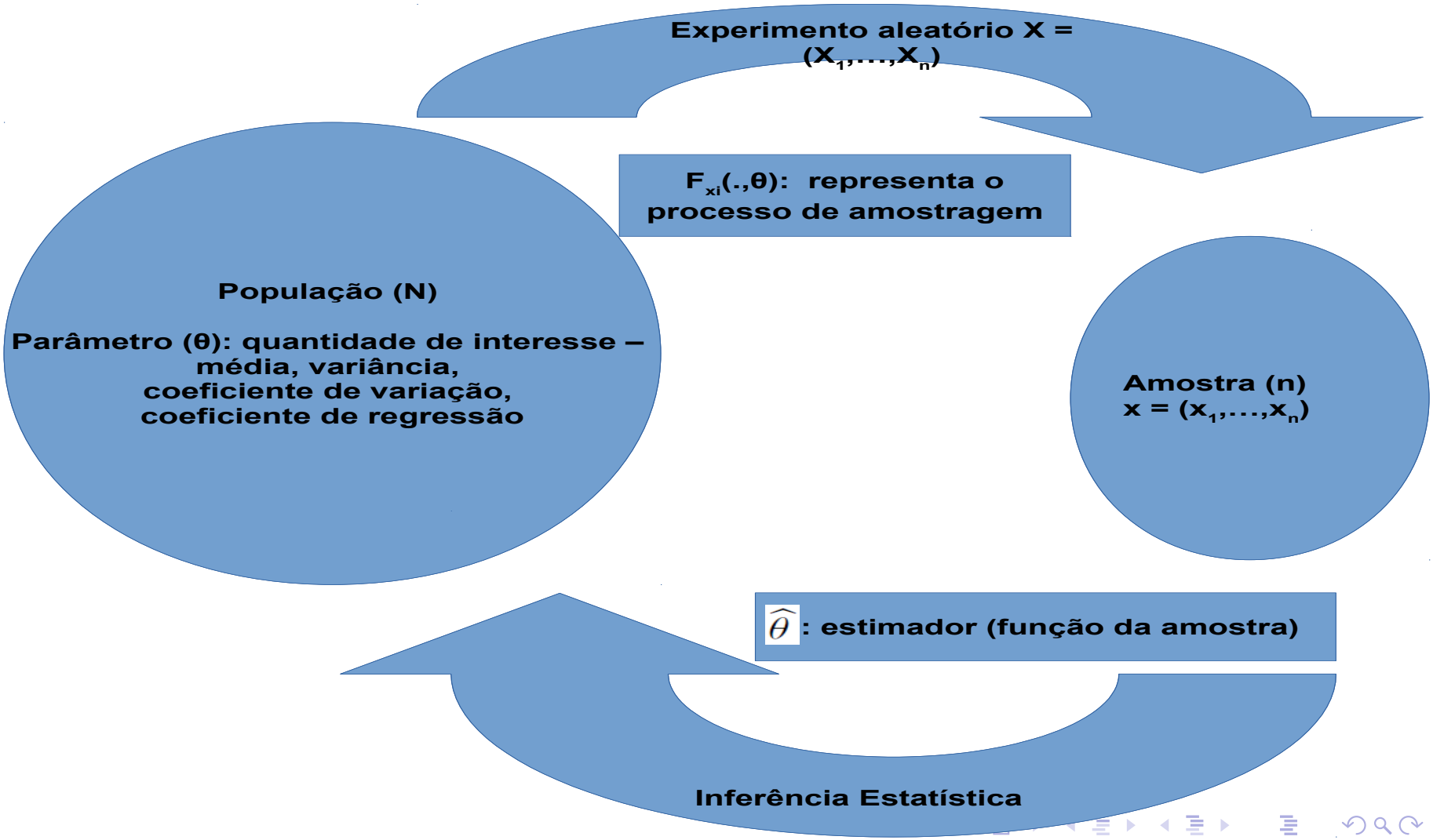
Problema Estatístico

- Considere um espaço estatístico $(\mathcal{X}, \mathcal{A}, \mathcal{P}_\theta)$, em que :
  - $\mathcal{X}$: suporte do modelo estatístico ou da família de distribuições de probabilidade $(\mathcal{P}_\theta)$ - $\mathcal{X} = \{\boldsymbol{x} \in \mathcal{R}^n, f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) > 0\}$.
  - $\mathcal{A}$: $\sigma-$álgebra (conjunto que contem todos os subconjuntos de (interesse) de $\mathcal{X}$).
  - θ: parâmetro de interesse (possivelmente um vetor, digamos: $\boldsymbol{\theta}$).
  - $\mathcal{P}_\theta$: família de medidas de probabilidade (família de distribuições de probabilidadade)- $\mathcal{P}_\theta : \mathcal{A} \to [0, 1], \forall \theta \in \Theta$, em que $\Theta \in \mathcal{R}^p$ é o espaço paramétrico (conjunto de posíveis valores que o parâmetro pode assumir).

Problema Estatístico

- Objetivo: inferir a respeito do valor (ou valores) mais provável(is) de θ com base em um experimento aleatório (envolvendo $\mathcal{P}_\theta$)
- O(s) modelo(s) estatístico(s) e o espaço amostral são induzidos pelo experimento aleatório, objetivos específicos e a(s) va(s) de interesse.
- Em geral temos diversas variáveis aleatórias envolvidas no experimento aleatório (entrevistas, amostragem, experimento científico etc).

Esquema representativo do problema estatístico



Estrutura geral

- Com base em uma amostra aleatória de tamanho n, digamos $\boldsymbol{X} = (X_1, ..., X_n)'$, obtida através de um processo de amostragem aleatória simples sem reposição (AAS) , deseja-se inferir a respeito do (verdadeiro) valor de (do parâmetro) θ. Ou seja, $X_i \overset{iid}{\sim} X$, $i = 1, 2, ..., n$, em que X é uma va de interesse (iid $\equiv$ independentes e identicamente distribuídas).
- Estimador $\widehat{\theta} = f(\boldsymbol{X})$ (va). Estimativa $\widetilde{\theta} = f(\boldsymbol{x})$ (número).
- O modelo (estatístico), $F_X(.; \theta)$ será assumido como conhecido em termos de sua forma paramétrica (normal, gama, uniforme, Poisson etc).

Estrutura geral

- Assim, teremos: $X_i \overset{iid}{\sim} F_X(.;\theta), i = 1, 2, ..., n$.
- Em geral, o parâmetro representa (ou está associado a) alguma quantidade de interesse da população (média, variância, assimetria, curtose).
- População: infinita ou infinitamente grande.

Comentários

- Na grande maioria das situações, os modelos estatísticos de interesse são (muito) mais complexos do que aqueles que serão visto ao longo do curso.
- Dificilmente, para eles é possível determinar os resultados (de otimilidade) e propriedades que serão vistas.
- Então, porque estudá-los...?

Respostas...

- Os conceitos de inferência são fundamentais para entender/estudar outras áreas da Estatística como: modelos de regressão, análise de sobrevivência, estatística não-paramétrica, análise multivariada etc.
- O conteúdo do curso pode ser adaptado, para tais situações. Em particular, pode-se buscar estimadores e testes de hipótese ótimos, porem, restritos a uma classe menor.
- Os métodos de construção de estimadores, intervalos de confiança (que, eventualmente, podem levar a resultados ótimos) e testes de hipótese podem ser adaptados (ainda que parcialmente) para algumas dessas situações (mesmo big/high dimensional data).

Exemplo 1: Distribuição de Bernoulli

- Considere que se quer estimar a proporção (θ) de indivíduos de uma determinada população, que tem sangue A^+. Assim, sob uma amostra de tamanho n, temos que:

$$X_i = \begin{cases} 1, & \text{se o i-ésimo elemento (indivíduo) da amostra for do tipo } A^+ \\ 0, & \text{caso contrário} \end{cases}$$

ou seja, $X_i \overset{iid}{\sim}$ Bernoulli(θ).

Exemplo 1: Distribuição de Bernoulli

- Para discussões sobre mecanismo de amostragem e análise sob estes, veja `http://www.ime.unicamp.br/~cnaber/Material_Amostragem_2S_2018.htm` e referências nele contidas.

  Assumiremos que esta etapa fora conduzida de forma apropriada.
- Neste caso $f_{X_i}(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i)$.
- Além disso, um bom (ótimo) estimador, em geral, é obtido reproduzindo-se na amostra, a definição (populacional) do parâmetro (ainda que a população seja infinita).

Exemplo 1: Distribuição de Bernoulli

- Assim $\widehat{\theta} \equiv \widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, é um bom candidato a estimador (ótimo).
- Usando-se apenas propriedades probabilísticas, temos que:
  - $\mathcal{E}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(X_i) = \theta\frac{n}{n} = \theta$ (considerando que cada esperança existe)
  - $\mathcal{V}(\widehat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathcal{V}(X_i) = \frac{\theta(1-\theta)}{n}$ (as variáveis são independentes)
  - $\widehat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$(Lei Fraca dos Grandes Números de Tchebyshev), assim, temos que $\widehat{\theta}$ é um estimador consistente.
  - $\mathcal{V}(\widehat{\theta}) = \frac{\theta(1-\theta)}{n} \xrightarrow[n\to\infty]{} 0$

Exemplo 1: Distribuição de Bernoulli

- Continuando....
  - $\frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow[n\to\infty]{D} N(0,1)$ (Teorema Central do Limite clássico).
  - Encontrar "limites de confiança": $P(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) = \gamma$, $\gamma \in (0,1)$.
  - Por outro lado, se utilizarmos $Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}$, é fácil encontrar (números), digamos q_1 e q_2,

$$P(q_1 \leq Z_n \leq q_2) \approx \gamma$$

Exemplo 1: Distribuição de Bernoulli

- Continuando....
  - Também podemos provar que: $W_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} \xrightarrow[n\to\infty]{D} N(0,1)$ (Exercício), mas a uma taxa menor (precisa de um tamanho de amostra maior), em relação à Z_n
  - Com os resultados deste slide e do anterior, podemos construir estimativas intervalares.

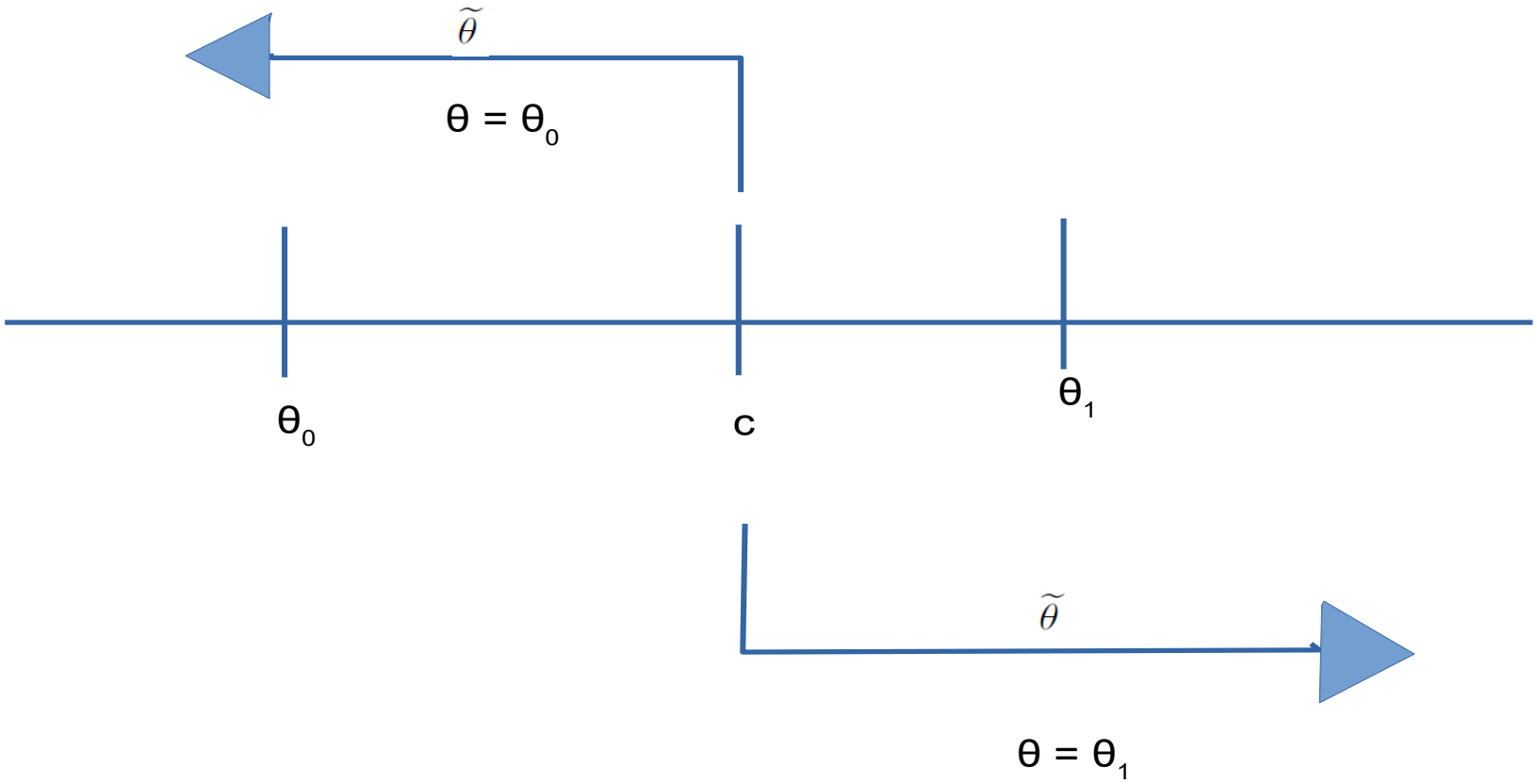
Exemplo 1: Distribuição de Bernoulli

- Suponha o interesse em testar um dos três seguintes conjuntos de hipóteses (estatísticas):
  - $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, $\Theta = \{\theta_0, \theta_1\}$, $\theta_0 < \theta_1$, $\theta_i \in [0, 1]$.
  - $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$, $\Theta = [\theta_0, 1]$, $\theta \in [0, 1]$.
  - $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, $\Theta = [0, 1]$, $\theta_0 \in [0, 1]$.

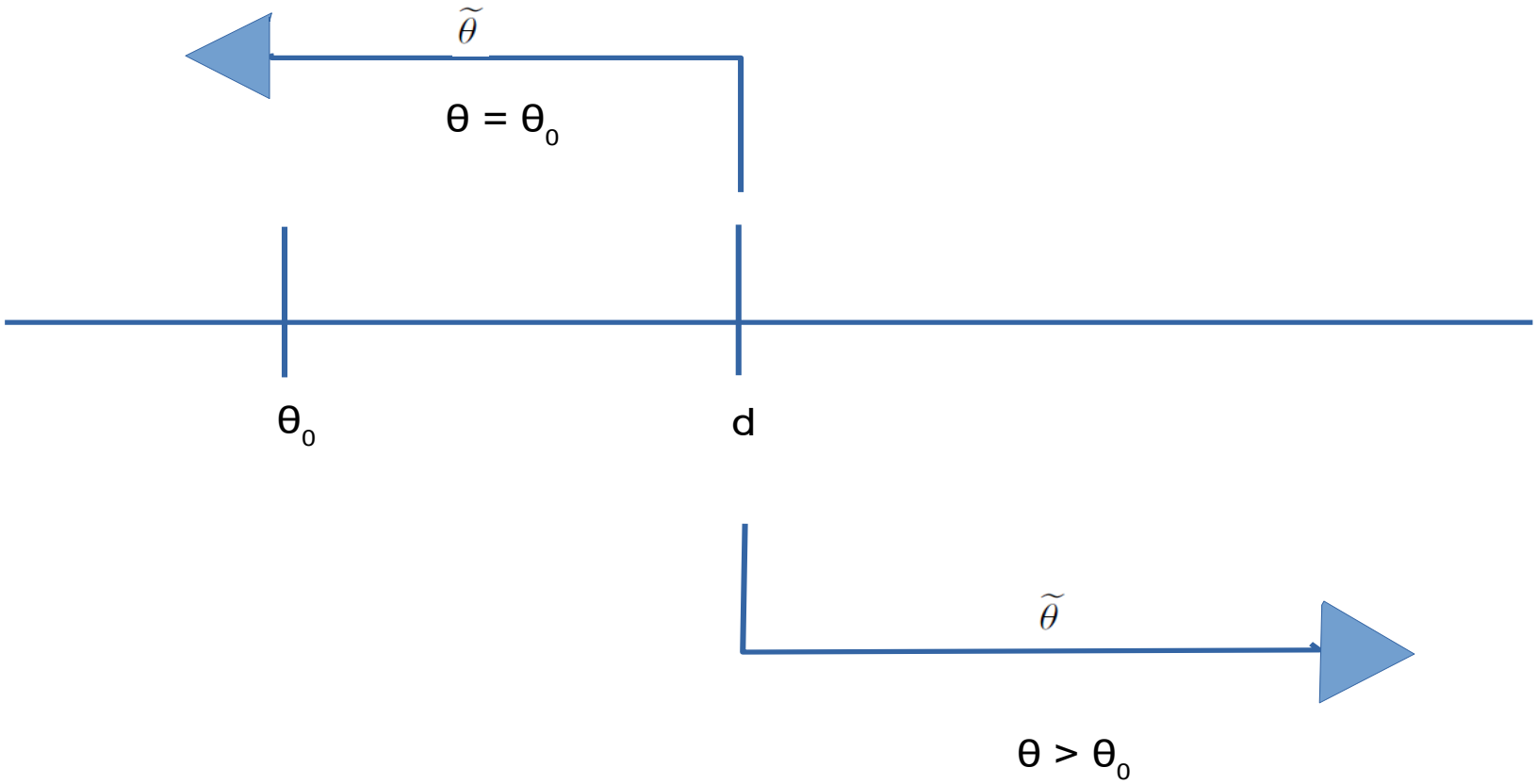
Exemplo 1: Distribuição de Bernoulli

- Com base em alguma função da amostra ($\widehat{\theta} = g(\boldsymbol{X})$), tem-se por objetivo estabelecer critérios para se avaliar a veracidade de cada hipótese. Por exemplo:
  - Se $\widetilde{\theta} > c$, conclui-se que $\theta = \theta_1$, caso contrário, conclui-se que $\theta = \theta_0$, em que $c \in [0, 1]$.
  - Se $\widetilde{\theta} > d$, conclui-se que $\theta > \theta_0$, caso contrário, conclui-se que $\theta = \theta_0$, em que $d \in [0, 1]$.
  - Se $\widetilde{\theta} > c_2$ ou $\widetilde{\theta} < c_1$, conclui-se que $\theta \neq \theta_0$, caso contrário, conclui-se que $\theta = \theta_0$, em que $c_i \in [0, 1], i = 1, 2$.

Hipóteses 1



Hipóteses 2



Hipóteses 3