INFERÊNCIA ESTATÍSTICA

ESTIMAÇÃO PARA A PROPORÇÃO POPULACIONAL p

Objetivo

Estimar uma proporção *p* (desconhecida) de elementos em uma população, apresentando certa característica de interesse, a partir da informação fornecida por uma amostra.

Exemplos:

p: proporção de alunos da USP que foram ao teatro pelo menos uma vez no último mês;

p: proporção de consumidores satisfeitos com os serviços prestados por uma empresa telefônica;

p: proporção de eleitores da cidade de São Paulo que votariam em um determinado candidato, caso a eleição para presidente se realizasse hoje;

p: proporção de crianças de 2 a 6 anos, do estado de São Paulo, que não estão matriculadas em escola de educação infantil.

Dois possíveis procedimentos de estimação:

- **Estimação pontual**
- **Estimação intervalar**

- Vamos observar *n* elementos, extraídos ao acaso e com reposição da população;

- Para cada elemento selecionado, verificamos a presença (sucesso) ou não (fracasso) da característica de interesse.

Estimador pontual

O ***estimador pontual para p***, também denominado ***proporção amostral***, é definido como

$$\hat{p} = \frac{X}{n},$$

sendo que

X denota o número de elementos na amostra que apresentam a característica;

n denota o tamanho da amostra coletada.

Se observamos o valor *k* da v. a. X, obtemos $\hat{p} = k / n$ que denominamos ***estimativa pontual para p.***

Exemplo 1: Sejam,

p: proporção de alunos da USP que foram ao teatro pelo menos uma vez no último mês, e

X: número de estudantes que respondem "sim" em uma pesquisa com *n* entrevistados.

Suponha que foram entrevistados $n = 500$ estudantes e que, desses, $k = 100$ teriam afirmado que foram ao teatro pelo menos uma vez no último mês.

A ***estimativa pontual*** (***proporção amostral***) ***para p*** é dada por:

$$\hat{p} = \frac{k}{n} = \frac{100}{500} = 0{,}20\,,$$

ou seja, 20% dos estudantes *entrevistados* afirmaram que foram ao teatro pelo menos uma vez no último mês.

Note que, outra amostra de mesmo tamanho pode levar a uma outra estimativa pontual para *p*.

Estimativa intervalar ou intervalo de confiança

•Para uma amostra observada, os estimadores pontuais fornecem como estimativa um <u>único valor numérico</u> para o parâmetro.

•Os estimadores pontuais são <u>variáveis aleatórias</u> e, portanto, possuem uma distribuição de probabilidade, em geral, denominada *distribuição amostral*.

Idéia: construir intervalos de confiança, que incorporem à estimativa pontual informações a respeito de sua variabilidade (erro amostral).

Intervalos de confiança são obtidos por meio da ***distribuição amostral do estimador pontual***.

A *estimativa intervalar* corresponde a um intervalo determinado da seguinte maneira:

$$\left[\hat{p} - \varepsilon ; \hat{p} + \varepsilon\right],$$

sendo ε o erro amostral ou margem de erro.

Pergunta: ***Como encontrar*** ε ***?***

Seja P(ε) a probabilidade da estimativa pontual estar a uma distância de, no máximo, ε da proporção verdadeira *p*, ou seja,

$$P(\varepsilon) = P\left(\left| \hat{p} - p \right| \leq \varepsilon \right).$$

A probabilidade P(ε) é também denominada **coeficiente de confiança do intervalo**, que denotamos pela letra grega γ (gama).

Afirma-se ainda que a estimativa intervalar tem coeficiente de confiança γ = P(ε).

Formalmente,

$$P(\varepsilon) = P(|\hat{p} - p| \leq \varepsilon) = P\left(\left|\frac{X}{n} - p\right| \leq \varepsilon\right)$$

$$= P\left(p - \varepsilon \leq \frac{X}{n} \leq p + \varepsilon\right) = P(np - n\varepsilon \leq X \leq np + n\varepsilon)$$

$$= P\left\{-\frac{n\varepsilon}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right\}.$$

Como $X \sim b(n,p)$ temos que, para n grande, a variável aleatória

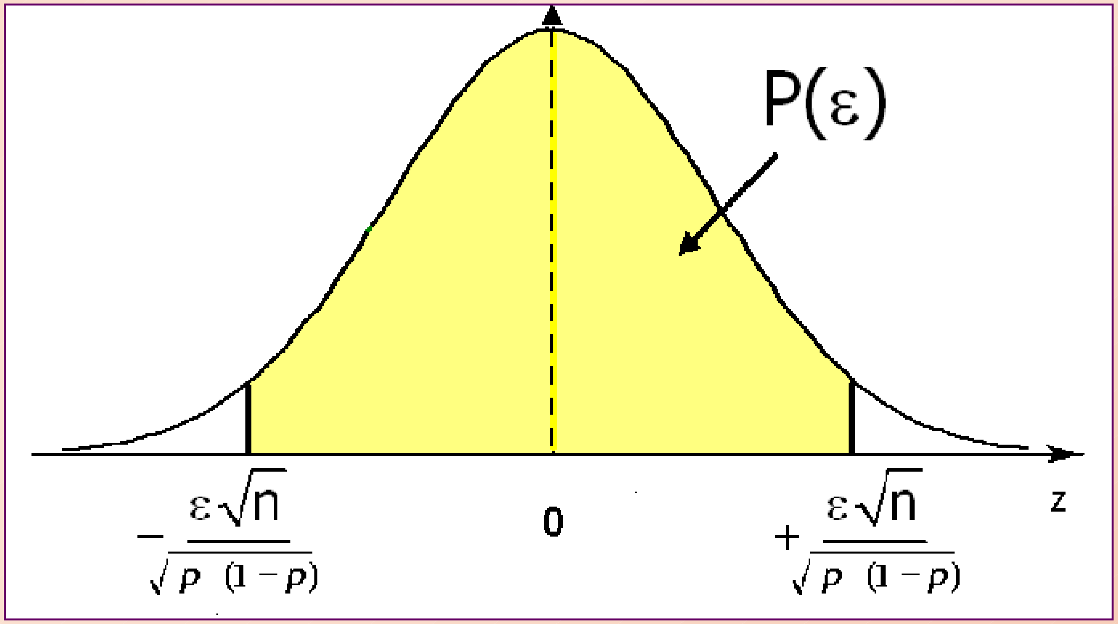
$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

tem distribuição N(0,1).

Deste modo, para *n* grande,

$$P(\varepsilon) \cong P\left\{-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right\},$$

onde Z ~ N(0,1).

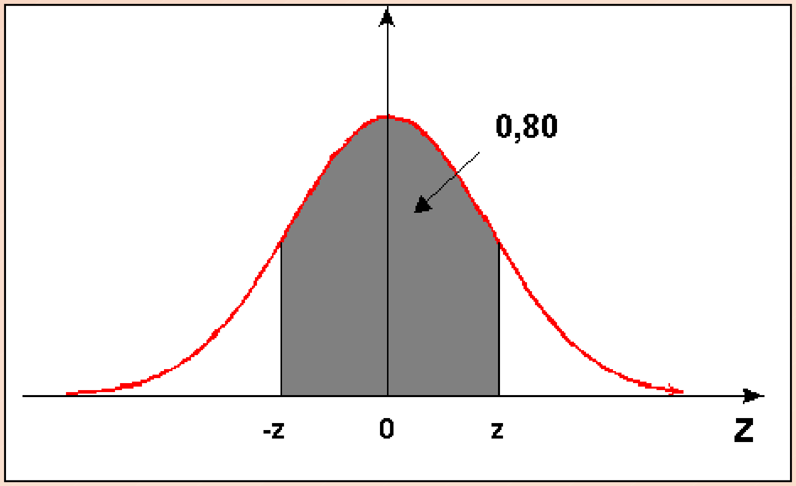


Denotando $\dfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} = z$, temos que

$P(\varepsilon) = \gamma = P(-z \le Z \le z)$.

Assim, podemos obter *z* conhecendo-se γ (ou $P(\varepsilon)$).

Por exemplo, considere $\gamma = 0{,}80$.



z é tal que $A(z) = 0{,}90$.

Pela tabela, temos $z = 1{,}28$.

Erro da estimativa intervalar

Da igualdade $\quad z = \dfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$,

é imediato mostrar que o **erro amostral ε** é dado por

$$\varepsilon = z\sqrt{\frac{p(1-p)}{n}} \;,$$

onde *z* é tal que $\gamma = P(-z \le Z \le z)$, com $Z \sim N(0,1)$.

Dimensionamento da amostra

Da relação $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$,

segue que o **tamanho amostral** ***n***, dados γ e a margem de erro ε, tem a forma

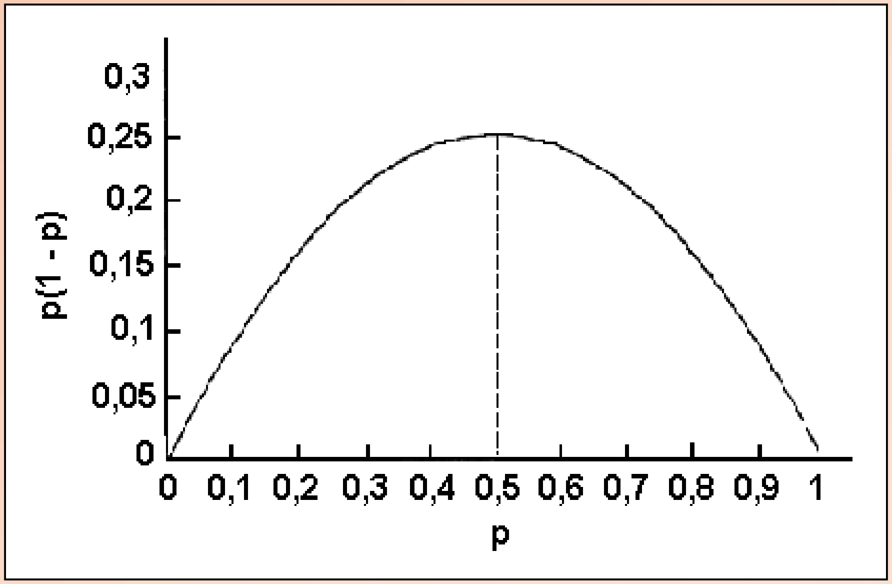
$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p),$$

onde z é tal que $\gamma = P(-z \leq Z \leq z)$ e $Z \sim N(0,1)$.

Entretanto, nesta expressão, *n* depende de *p(1-p)*, que é desconhecido.

- ***Como calcular o valor de n?***

Gráfico da função *p(1-p)*, para $0 \leq p \leq 1$.



Pela figura observamos que:

- **a função *p(1-p)* é uma parábola simétrica em torno de $p = 0,5$;**
- **o máximo de *p(1-p)* é 0,25, alcançado quando $p = 0,5$.**

Assim, na prática, substituímos *p(1-p)* por seu valor máximo, obtendo

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}25\,,$$

que pode fornecer um valor de *n* maior do que o necessário.

Exemplo 2:

No exemplo da USP (Exemplo 1) suponha que nenhuma amostra foi coletada. Quantos estudantes precisamos consultar de modo que a estimativa pontual esteja, no máximo, a 0,02 da proporção verdadeira *p*, com uma probabilidade de 0,95?

Dados do problema:

ε = 0,02 (erro da estimativa);

$P(\varepsilon) = \gamma = 0{,}95 \;\; \Rightarrow \;\; z = 1{,}96.$

$$n = \left(\frac{1{,}96}{0{,}02}\right)^2 p(1-p) \leq \left(\frac{1{,}96}{0{,}02}\right)^2 0{,}25 = 2401 \text{ estudantes} .$$

Pergunta: ***É possível reduzir o tamanho da amostra quando temos alguma informação a respeito de p?***

Por exemplo, sabemos que:

- ***p*** **não é superior a 0,30, ou**
- ***p*** **é pelo menos 0,80, ou**
- ***p*** **está entre 0,30 e 0,60.**

Resposta: ***Depende do tipo de informação sobre p.***

Em alguns casos, podemos substituir a informação *p(1-p)*, que aparece na expressão de *n*, por um valor menor que 0,25.

Redução do tamanho da amostra

Vimos que, se nada sabemos sobre o valor de *p*, no cálculo de *n*, substituímos *p(1-p)* por seu valor máximo, e calculamos

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}25\,.$$

Se temos a informação de que *p é no máximo 0,30* ($p \leq 0{,}30$), então o valor máximo de *p(1-p)* será dado por 0,3x0,7 = 0,21.

Logo, reduzimos o valor de *n* para

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}21.$$

Agora, se *p é pelo menos 0,80* (p ≥ 0,80), então o máximo de *p(1-p)* é 0,8 x 0,2 = 0,16 e temos

$$\mathbf{n} = \left(\frac{\mathbf{z}}{\varepsilon}\right)^2 \mathbf{0{,}16}\,.$$

Mas, se *0,30 ≤ p ≤ 0,60*, o máximo de *p(1-p)* é 0,5x0,5 = 0,25 e, neste caso, não há redução, ou seja,

$$\mathbf{n} = \left(\frac{\mathbf{z}}{\varepsilon}\right)^2 \mathbf{0{,}25}.$$

Exemplo 3:

No Exemplo 2, suponha que temos a informação de que no máximo 30% dos alunos da USP foram ao teatro no último mês. Portanto, temos que

$p \leq 0{,}30$ e, como vimos, o máximo de $p(1-p)$ neste caso é 0,21.

Assim, precisamos amostrar

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}21 = \left(\frac{1{,}96}{0{,}02}\right)^2 0{,}21 = 2017 \text{ estudantes },$$

conseguindo uma redução de 2401 – 2017 = 384 estudantes.

Intervalo de confiança para *p*

Vimos que a estimativa intervalar para ***p*** tem a forma: $\left[\hat{p} - \varepsilon\,;\, \hat{p} + \varepsilon\right]$,

com $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$ e *z* tal que γ = P(-z ≤ Z ≤ z) na N(0,1).

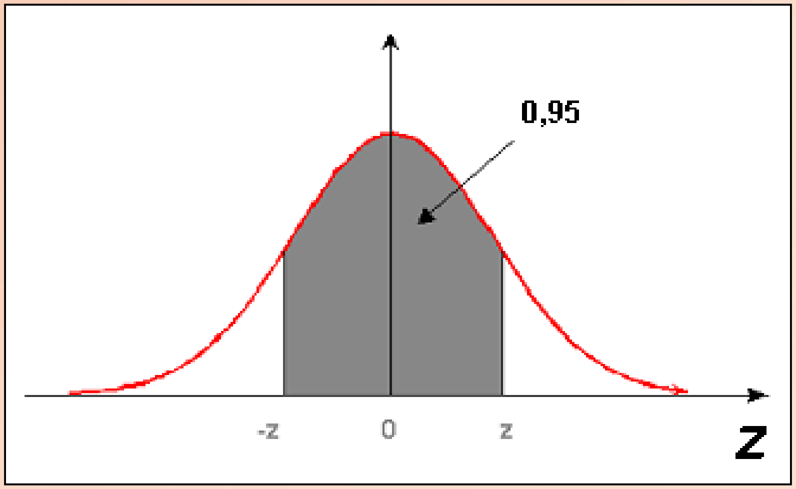
Na prática, substituímos a proporção desconhecida p pela proporção amostral $\hat{p}$***, obtendo o seguinte intervalo de confiança com coeficiente de confiança*** γ ***:***

$$IC(p\,;\gamma) = \left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\;\;;\;\; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exemplo 4:

No exemplo da USP, temos n = 500 e $\hat{p}$ = 0,20.

Construir um intervalo de confiança para *p* com coeficiente de confiança γ = 0,95.



Como γ = 0,95 fornece *z* = 1,96, o intervalo é dado por:

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;\; ; \;\; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$= \left[0{,}20 - 1{,}96\sqrt{\frac{0{,}20 \times 0{,}80}{500}} \;\; ; \;\; 0{,}20 + 1{,}96\sqrt{\frac{0{,}20 \times 0{,}80}{500}}\right]$$

$$= [\,0{,}20 - 0{,}035\, ;\; 0{,}20 + 0{,}035\,] = [\,0{,}165\, ;\; 0{,}235\,].$$

Nesse intervalo (γ=0,95), a estimativa pontual para *p* é 0,20, com um erro amostral ε igual a 0,035.

Interpretação do IC com $\gamma = 95\%$:

Se sortearmos 100 amostras de tamanho n=500 e construirmos os respectivos 100 intervalos de confiança, com coeficiente de confiança de 95%, esperamos que, aproximadamente, 95 destes intervalos contenham o verdadeiro valor de *p*.

Comentários:

Da expressão $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$, é possível concluir que:

- para γ fixado, o erro diminui com o aumento de *n*.
- para *n* fixado, o erro aumenta com o aumento de γ.

NOÇÕES DE TESTE DE HIPÓTESES (I)

Teste de hipóteses para a proporção populacional

Estimação	Teste de Hipóteses
Qual é a probabilidade de "cara"no lançamento de uma moeda?	A moeda é honesta ou é desequilibrada?
Qual é a proporção de votos que o candidato A tem nas eleições?	O candidato A vencerá as eleições ?
Qual é a proporção de motoristas que tiveram sua carteira apreendida após a vigência da nova lei de trânsito?	Pelo menos 2% dos motoristas habilitados de SP tiveram suas carteiras apreendidas após a entrada da nova lei do trânsito ou não?

Introdução

Em **estimação** o objetivo é “estimar” o valor desconhecido da proporção p de “indivíduos” em uma população com determinada característica.

A estimativa é baseada no número X de “indivíduos” com a característica numa amostra casual simples de tamanho n.

Entretanto, se o objetivo for saber se o valor observado x nessa amostra, dá ou não suporte a uma conjectura sobre o valor de p, trata-se de **testar hipóteses.**

Exemplo 1: Queremos avaliar se uma moeda é honesta.

Ou seja, queremos testar a

hipótese nula *H*: a moeda é honesta

contra a

hipótese alternativa *A*: a moeda não é honesta

Em linguagem estatística, essas hipóteses podem ser reescritas como:

$$H: p = 0{,}5$$

$$A: p \neq 0{,}5$$

com p sendo a probabilidade de “cara” da moeda.

Hipóteses

De uma maneira geral, uma **hipótese estatística** é uma afirmação ou conjectura sobre um parâmetro da distribuição de uma variável aleatória.

Hipótese nula: afirmação ou conjectura sobre p contra a qual estaremos buscando evidência nos dados amostrais.

Hipótese alternativa: afirmação ou conjectura sobre p que esperamos ser verdadeira.

No nosso exemplo, se considerarmos 12 lançamentos independentes da moeda e denotarmos por X o número de caras nesses lançamentos, então o parâmetro é a proporção de caras p e

$$X \sim \textbf{binomial}\ (12;\ p\ )$$

Se observarmos 5 caras em 12 lançamentos independentes da moeda, o que podemos concluir?

E se observarmos 4 caras? Ou 10 caras?
Ou 12 caras?

Podemos considerar uma **regra de decisão**, como por exemplo,

> "Se, em 12 lançamentos da moeda, observarmos 0,1, 2, 3, 9, 10, 11 ou 12 caras, então rejeitamos a hipótese nula H de que a moeda é honesta;
>
> caso contrário, aceitamos a hipótese H."

Testar uma hipótese estatística é estabelecer uma regra que nos permita, com base na informação de uma amostra, decidir pela rejeição ou não de *H*.

No exemplo, o conjunto de valores de X que levam à rejeição da hipótese nula *H* é {0, 1, 2, 3, 9, 10, 11, 12}, o qual denominamos de **região crítica (RC)** ou **região de rejeição de *H***, ou seja,

RC = {0, 1, 2, 3, 9 , 10, 11, 12} : **região crítica**

RC^c = {4, 5, 6, 7, 8} : **região de aceitação de *H***

Regra de decisão (teste):

seja x o valor observado na amostra da variável X, então

$x \in$ RC $\Rightarrow$ rejeitamos *H*

$x \notin$ RC $\Rightarrow$ não rejeitamos *H*

No exemplo da moeda, suponha que observamos 4 caras, isto é, $x = 4$.

Como **4 $\notin$ RC $\Rightarrow$ não rejeitamos *H*** (não temos evidência suficiente de que a moeda seja desequilibrada).

Será que nossa conclusão está correta?

Ao decidir pela rejeição ou não da hipótese nula *H*, podemos cometer *dois tipos de erro*.

Erros

Erro tipo I: Rejeitar *H* quando *H* é verdadeira
(afirmar que uma moeda não é honesta quando, na verdade, ela é).

Erro tipo II: Não rejeitar *H* quando *H* é falsa
(afirmar que uma moeda é honesta quando, na verdade, ela é desequilibrada).

Probabilidades de erros

P(erro I) = P(rejeitar H | H é verdadeira**) =** α

α **: nível de significância do teste**

P(erro II) = P(não rejeitar H | H é falsa**) =** β

1 - β **: poder do teste**

No exemplo da moeda,

$$RC = \{0,1,2,3,9,10,11,12\}$$

α = P(erro I) = P(rejeitar H | H verdadeira)

$= P(X \in \mathbf{RC} \mid p=0,5)$

$= P(X=0 \mid p=0,5) + \ldots + P(X=3 \mid p=0,5) + P(X=9 \mid p=0,5) +$

$\ldots + P(X=12 \mid p=0,5)$ $\Rightarrow$

$= 0,000244 + 0,00293 + 0,016113 + 0,053711 + 0,053711 +$

$0,016113 + 0,00293 + 0,000244$

$=$ **0,1460**

Decisão	Verdadeiro valor de *p*	
	p **= 0,5** (*H* é verd.)	$p \neq$ **0,5** (*A* é verd.)
Não rejeitar *H*	Decisão correta **1 -** $\alpha =$ 0,8540	**Erro II** β
Rejeitar *H*	**Erro I** $\alpha =$ 0,1460	Decisão correta **1 -** β

Se alterarmos a regra de decisão para **RC** = {0, 1, 2, 10, 11, 12}, isto é, concluiremos que a moeda é desonesta se o número de caras for 0, 1, 2, 10, 11 ou 12, o que acontece com o nível de significância do teste α (probabilidade de erro tipo I)?

α = P(erro I) = P(rejeitar H | H verdadeira)=P($X \in$ **RC** |p=0,5)

= P(X=0 | p=0,5) + ... + P(X=2 | p=0,5) + P(X=10 | p=0,5)+

... + P(X=12 | p=0,5) $\Rightarrow$

= 0,000244 + 0,00293 + 0,016113 + 0,016113 + 0,00293 + 0,000244

= **0,0384**

Regiões críticas e níveis de significância α (Exemplo 1 – moeda)

RC	α
{0, 1, 2, 3, 9, 10, 11, 12}	0,1460
{0, 1, 2, 10, 11, 12}	0,0384
{0, 1, 11, 12}	0,0063

Os valores de nível de significância α usualmente adotados são entre 1% e 10%.

Até agora, o procedimento foi

escolher RC $\Rightarrow$ **determinar** α

Alternativamente, podemos

fixar α $\Rightarrow$ **determinar RC**

Determinação da região crítica

Exemplo 2: Suponha que um medicamento existente no mercado produza o efeito desejado em 60% dos casos nos quais o mesmo é aplicado.

Um laboratório produz um novo medicamento e afirma que ele é melhor do que o existente.

Objetivo: Verificar estatisticamente se a afirmação do laboratório é verdadeira.

Aplicou-se o medicamento em $n = 10$ pacientes.

Sendo X o nº de pacientes, dentre os 10, para os quais o novo medicamento produz o efeito desejado, temos que,

$$X \sim \mathbf{b}\,(\mathbf{10};\, p),$$

com p sendo a proporção de pacientes para os quais o novo medicamento é eficaz.

(1) Hipóteses estatísticas:

$$\boxed{\begin{array}{l} H\colon p = 0{,}6 \\ A\colon p > 0{,}6 \end{array}}$$

que correspondem a

H: o novo medicamento é similar ao existente

A: o novo medicamento é melhor, mais efetivo

(2) Fixemos o nível de significância em 5% ($\alpha = 0{,}05$).

(3) A região crítica deve ter a forma:

$$\mathbf{RC} = \{ X \geq k \}$$

O valor de k deve ser tal que

P(erro I) = $P(X \in \mathbf{RC} \mid p = 0{,}6) = P(X \geq k) = \alpha$,

Pela tabela da **binomial** (10; 0,6), ⇒

para $k = 9$, $P(X \geq 9) = 0{,}0463$

para $k = 8$, $P(X \geq 8) = 0{,}1672$

Portanto, $\mathbf{RC} = \{X \geq 9\}$, garante um erro tipo I de no máximo 5% (na realidade, $\alpha =$ **0,0463**).

Hipóteses alternativas unilaterais e bilaterais

No exemplo 2 as hipóteses nula e alternativa são:

$$H: p = 0{,}6 \quad \text{e} \quad A: p > 0{,}6$$

isto é, desejamos detectar desvios em p apenas em uma direção, ou seja, desvios à "direita" de 0,6.

Neste caso, dizemos que a hipótese alternativa é **unilateral**.

No exemplo 1 (da moeda), como as hipóteses são

$$H: p = 0{,}5 \quad \text{e} \quad A: p \neq 0{,}5$$

dizemos que a hipótese alternativa é **bilateral** (detectariamos desvios em torno de $p = 0{,}5$ em qualquer direção).

Exemplo 3: A proporção de analfabetos em um município era de 15% na gestão anterior. O prefeito atual implantou um programa de alfabetização desde o início de sua gestão e afirma que após 2 anos reduziu a proporção de analfabetos.

Para verificar a afirmação do prefeito 60 cidadãos foram entrevistados.

Seja X o número de analfabetos entre 60 cidadãos entrevistados. Então,

$$X \sim \mathbf{bin}(60; p),$$

sendo p a proporção atual de analfabetos (após o programa de alfabetização).

(1) As **hipóteses de interesse** são

H: a proporção de analfabetos não se alterou (a afirmação do prefeito está incorreta).

A: a proporção de analfabetos diminuiu (afirmação do prefeito está correta).

Equivalentemente,

$$\boxed{\begin{array}{l} H\text{: } p = 0{,}15 \\ A\text{: } p < 0{,}15 \end{array}}$$

(2) Vamos **fixar α = 0,05**.

(3) A **região crítica** deve ter a forma:

$$\textbf{RC} = \{ X \leq k \}$$

O valor de k deve ser tal que P(erro I) = α, ou seja,

$$P(X \leq k \mid p = 0{,}15) = 0{,}05.$$

Pela tabela da **binomial**(60; 0,15), $\Rightarrow$

$$\boxed{\text{RC} = \{ X \leq 4\}}$$

Na realidade temos $\alpha = 0{,}0424$.

(4) Buscar a **evidência na amostra** para concluir:

Se observamos 6 analfabetos entre os 60 entrevistados, qual é a conclusão?

(5) **Decisão e conclusão**

6 $\notin$ **RC** $\Rightarrow$ decidimos por não rejeitar H, ao nível de significância de 4,24%.

Concluímos que não temos evidência suficiente para afirmar que a proporção de analfabetos (após o programa de alfabetização) é inferior a 15%, isto é, não há evidência suficiente de que a afirmação do prefeito seja correta.

Resumo

(1) Estabelecer as **hipóteses**:

H: $p = p_0$ contra uma das alternativas

A: $p \neq p_0$, A: $p > p_0$ ou A: $p < p_0$.

(2) Escolher um **nível de significância** α.

(3) Determinar a **região crítica RC** da forma

$\{ X \leq k_1 , X \geq k_2 \}$, $\{ X \geq k \}$ ou $\{ X \leq k \}$,

respectivamente às hipóteses alternativas.

(4) Selecionar uma **amostra** casual simples e determinar o número x de “indivíduos” na amostra portadores do atributo desejado.

(5) **Decidir**, usando a evidência x, ao nível de significância α, e **concluir**.

Se $x \in$ RC $\Rightarrow$ rejeitamos H

$x \notin$ RC $\Rightarrow$ não rejeitamos H