EXPOSIÇÃO SOBRE O TEMA

Estimação Intervalar de Proporção

das disciplinas "Noções de Estatística" MAE0110 e MAE0116
na Universidade de São Paulo

Preparado por Prof. Dr. Vladimir Belitsky, docente do IME-USP
04 de abril de 2020

O material que completa a presente exposição está na página web
www.ime.usp.br/~belitsky

Há ma urna com bolas idênticas no tato, mas pintadas de duas cores diferentes: branca e preta. Denoto por p a proporção de bolas pretas. p é desconhecido e desejamos estimá-lo com base no resultado de amostra de n bolas retiradas da urna ao acaso e com reposição. Aqui n não é um valor desconhecido. É uma notação genérica para o valor a estabelicido antes de fazer a amostra. A notação genérica para o resultado de amostra é

$$x_1, x_2, \ldots, x_n \tag{1}$$

onde x_i representa a cor da i-ésima bola retirada. A representação pode usar qualquer codificação cômoda, mas, com a vista nas necessidades e comodidades dos argumentos futuros, será adotada a seguinte codificação específica

$$x_i = \begin{cases} 1, & \text{caso a } i\text{-ísima bola da amostra for preta} \\ 0, & \text{caso a } i\text{-ísima bola da amostra for branca} \end{cases} \tag{2}$$

Abaixo está um exemplo da amostra que pode surgir como resultado de retirada de $n = 20$ bolas:

$$\bullet \ \circ \ \bullet \ \bullet \ \bullet \ \circ \ \bullet \ \circ \ \circ \ \bullet \ \circ \ \bullet \ \circ \ \circ \ \bullet \ \bullet \ \bullet \ \circ \ \bullet \ \circ \quad (3)$$

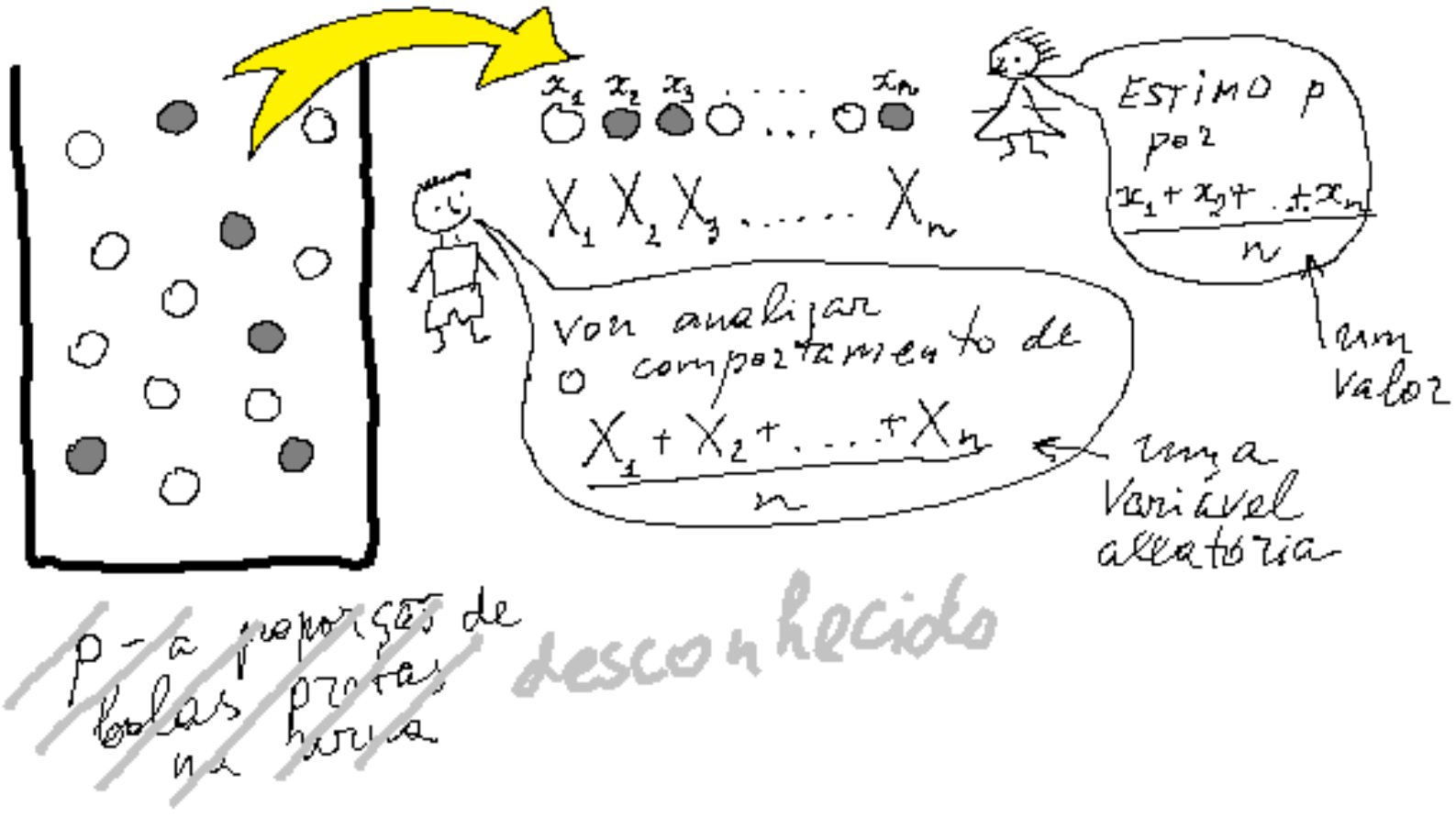
Os valores da sequencia x que representa essa amostra são assim:

$$1,\ 0,\ 1,\ 1,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 1,\ 1,\ 0,\ 1,\ 0 \quad (4)$$

As notações a serem usadas: n - o tamanho de amostra, k - a quantidade de bolas pretas na amostra: No caso do exemplo acima apresentado, $n = 20$, $k = 11$.
Com isso, $\frac{k}{n}$ é a proporção amostral de bolas pretas. É natural que a proporção amostral seja tomada como a estimativa pontual para a proporção populacional, quer dizer, para p. Vamos aceitar tal estimação sem dar amiores explicações e justificativas.

Nosso objetivo atual é analizar qual longe o valor estimado (p) pode estar em relação a sua estimativa

$$\frac{k}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{5}$$



Na perspectiva de "vai acontecer", tem-se que

$$X_i = \begin{cases} 1, \text{ com probabilidade } p \\ 0, \text{ com probabilidade } 1-p \end{cases}$$

e que

$$X_1, X_2, \ldots, X_n \text{ são independentes entre si}$$

Consequentemente

$$S = (X_1 + X_2 + \cdots + X_n) \sim Bin(n, p) \tag{6}$$

(embora não sabemos o valor de p, podemos alegar (6)).

Vamos usar as propriedades reveladas na transparência anterior para calcular a probabilidade do intervalo

$$\left[\frac{S}{n} - \varepsilon, \frac{S}{n} + \varepsilon\right] \tag{7}$$

conter p; aqui, ε é um parâmetro.

Então, o objetivo atual é calcular

$$\mathbf{P}\left\{p \in \left[\frac{S}{n} - \varepsilon, \frac{S}{n} + \varepsilon\right]\right\}$$

Recordo: $S = (X_1 + X_2 + \cdots + X_n)$.

Em conta abaixo usaremos que $a - \varepsilon \leq b \leq a + \varepsilon$ se e somente se $b - \varepsilon \leq a \leq b + \varepsilon$, para quaisquer $a, b \in \mathbb{R}$.

$$\begin{aligned} & \mathrm{P}\left\{\frac{S}{n} - \varepsilon \leq p \leq \frac{S}{n} + \varepsilon\right\} \\ = \; & \mathrm{P}\left\{p - \varepsilon \leq \frac{S}{n} \leq p + \varepsilon\right\} \\ = \; & \mathrm{P}\left\{n(p - \varepsilon) \leq S \leq n(p + \varepsilon)\right\} \end{aligned}$$

Substituo S por $Y \sim \mathcal{N}\left(np, np(1-p)\right)$ (é aqui que usamos fato que $X \sim Bin(n, p)$), e depois faço a padronização:

$$\begin{aligned} \approx \; & \mathrm{P}\left\{n(p - \varepsilon) \leq Y \leq n(p + \varepsilon)\right\} \\ = \; & \mathrm{P}\left\{-\frac{n\varepsilon}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right\} \\ = \; & \mathrm{P}\left\{-\frac{n\varepsilon}{\sqrt{np(1-p)}} \leq Z \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right\} \end{aligned}$$

Interpretação (do ponto de vista de "antes de fazer amostra"): o intervalo aleatório

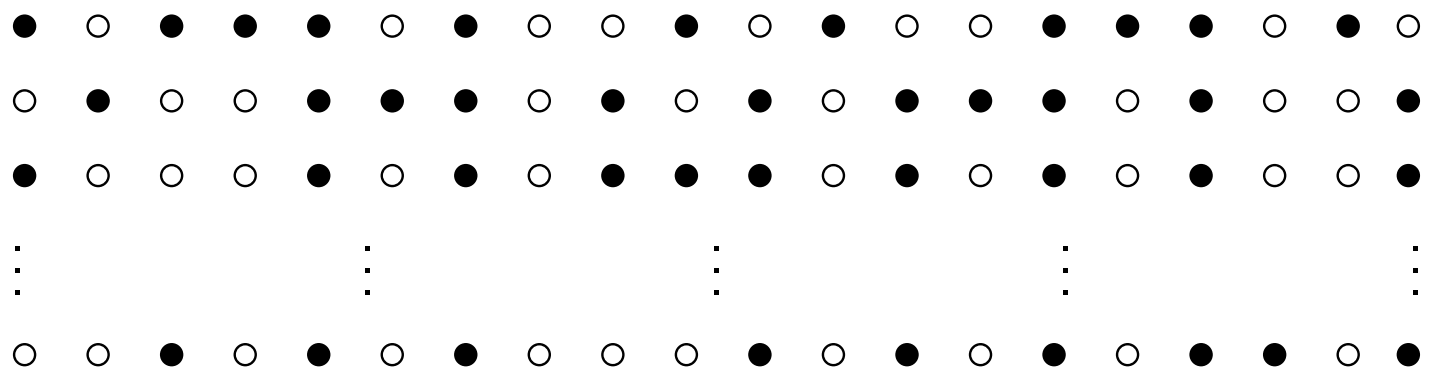
$$\left[\frac{S}{n} - \varepsilon, \frac{S}{n} + \varepsilon\right]$$

acertará o verdadeiro valor de p com

$$\mathrm{P}\left\{-\frac{n\varepsilon}{\sqrt{np(1-p)}} \leq Z \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right\}$$

isto é, com

$$\mathrm{P}\left\{-\frac{\varepsilon}{\sqrt{\left(\frac{p(1-p)}{n}\right)}} \leq Z \leq \frac{\varepsilon}{\sqrt{\left(\frac{p(1-p)}{n}\right)}}\right\}$$



Cada uma acarreta intervalo $[\frac{k}{20} - 0,15,\ \ \frac{k}{20} + 0,15]$ (com "seu" valor de k). Tais intervalos podem ser separados por bons (os que acertaram p) e ruins (os que não acertaram).

A quantidade de intervalos no saco ruim é

$$1 - \mathrm{P}\left\{ -\frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \leq Z \leq \frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \right\}$$

enquanto que a quantidade de intervalos no saco bom é

$$\mathrm{P}\left\{ -\frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \leq Z \leq \frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \right\}$$

Tem-se uma amostra só:

● ○ ● ● ● ○ ● ○ ○ ● ○ ● ○ ○ ● ● ● ○ ● ○

para a qual $k = 11$, e portanto, o intervalo correspondente (recordo, $\varepsilon = 0,15$) é

$$\left[\frac{11}{20} - 0,15,\ \frac{11}{20} + 0,15\right] = [0,4,\ 0,7]$$

o que pode-se dizer acerca dele "acertar" o verdadeiro valor de p?

A probabilidade dele não acertar é a probabilidade dele ter vindo do saco de intervalos ruins, que é

$$1 - \mathrm{P}\left\{ -\frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \leq Z \leq \frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \right\}$$

enquanto que a probabilidade dele acertar é a probabilidade dele ter vindo do saco de intervalos bons, que é

$$\mathrm{P}\left\{ -\frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \leq Z \leq \frac{0,15}{\sqrt{\left(\frac{p(1-p)}{20}\right)}} \right\}$$

Vamos chamar a segunda probabilidade de **coeficiente de confiança** que seu intervalo possui.

A resposta é bonita, mas depende de p, cujo valor é desconhecido (recorde o objetivo final de tudo que estamos fazendo é achar uma maneira boa de estimar p e, se possível quantificar sua bondade).

Pode colocar $\frac{k}{n}$, determinado pela sua amostra, no lugar do desconhecido p; há cálculo que mostra que tal substituição não afeta significativamente o resultado final. Isto é, o valor do coeficiente de confiança do itervalo está expressado por

$$\mathrm{P}\left\{ -\frac{0,15}{\sqrt{\left(\frac{\frac{11}{20}(1-\frac{11}{20})}{20}\right)}} \leq Z \leq \frac{0,15}{\sqrt{\left(\frac{\frac{11}{20}(1-\frac{11}{20})}{20}\right)}} \right\}$$

O intervalo de estimação tem o formato $\left[\frac{k}{n} - \varepsilon\,,\ \frac{k}{n} + \varepsilon\right]$

A cada intervalo de estimação atribui-se seu **coeficiente de confiança** que interpreta-se pela probabilidade do intervalo captar p;

Coeficiente de confiança será denotado por γ.

γ e ε vinculam-se um a outro via a seguinte fórmula (abaixo $Z \sim \mathcal{N}(0,1)$):

$$\underbrace{\varepsilon = z\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}}_{\text{primeira parte}},\ \text{onde}\ \underbrace{z > 0 \text{ é tal que } \gamma = \mathrm{P}\left[-z \leq Z \leq z\right]}_{\text{segunda parte}},$$

a fórmula vale para valores suficientemente grandes de n;

assumiremos implicitamente tal qualidade de n.

São os seguintes três tipos de problemas que você deve aprender a resolver:
Quando a amostra está conhecida (n e k estão conhecidos), há problema de clacular o coeficiente de confiança para um dado valor de margem de erro, e há outro problema que é calcular a margem de erro para um dado coeficiente de confiança.
As fórmulas são assim:

dado ε calcule $z = \dfrac{\varepsilon}{\sqrt{\left(\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}\right)}}$ e obtenha $\gamma = \mathrm{P}\left[-z \leq Z \leq z\right]$

dado γ ache z tal que $\gamma = \mathrm{P}\left[-z \leq Z \leq z\right]$ e obtenha

$$\varepsilon = z\sqrt{\left(\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}\right)}$$

O terceiro tipo aplica-se a situação quando a amostra ainda não foi feita. Nesse caso o problema é achar o valor mínimo de n que garante que quando formos fazer amostra de tamanho n, então, com o resultado dessa amostra, poderemos construir intervalo de confiança cujo coeficiente de confiança é de no mínimo um dado valor γ e cuja margem de erro é de no máximo um dado valor ε.

O tamanho mínimo de amostra (n_{min}) a ser feita para garantir que para qualquer resultado (k), a desconhecida proporção populacional (p) possa ser estimada com a margem de eror de no máximo ε e o coeficiente de confiança de no mínimo γ fixados antemão, determina-se pelas seguintes fórmulas:

$$n_{\text{min}} = M_{\mathcal{D}} \left(\frac{z}{\varepsilon}\right)^2 \quad \text{caso for conhecido apriori que } p \tag{8}$$

pode estar somente no conjunto $\mathcal{D}$; neste caso

$$M_{\mathcal{D}} = \max_{x \in \mathcal{D}} x(1-x); \tag{9}$$

em particular, se nenhuma informação acerca de p estar disponível, então $\mathcal{D} = [0, 1]$, $\max_{x \in [0,1]} x(1-x) = 0,5(1-0,5)$, e consequentemente

$$n_{\min} = 0,5(1-0,5)\left(\frac{z}{\varepsilon}\right)^2 = 0,25\left(\frac{z}{\varepsilon}\right)^2 \tag{10}$$

Em todas as fórmulas, z é um número positivo tal que $\gamma = \mathrm{P}[-z \leq Z \leq z]$, com $Z \sim \mathcal{N}(0, 1)$.

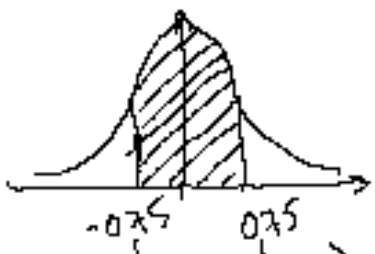
Suponha que decidi por $n = 100$. Ao fazer meu experimento, observei 20 bolas pretas. Suponha que decidi por $\varepsilon = 0,03$. Então, o *intervalo de confiança* resultante é

$$[0,2 - 0,03\,;\, 0,2 + 0,03]$$

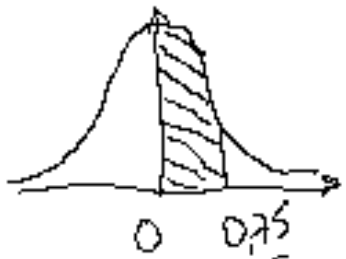
e o correspondente *coeficiente de confinaça* é

$$\begin{aligned}
&\mathrm{P}\left\{-\frac{100 \cdot 0,03}{\sqrt{100 \cdot 0,2(1-0,2)}} \leq Z \leq \frac{100 \cdot 0,03}{\sqrt{100 \cdot 0,2(1-0,2)}}\right\} \\
&= \mathrm{P}\left\{-0,75 \leq Z \leq 0,75\right\} = 2\left(A(0,75) - 0,5\right) \\
&= 2(0,7734 - 0,5) = 0,5468 \approx 0,55
\end{aligned}$$

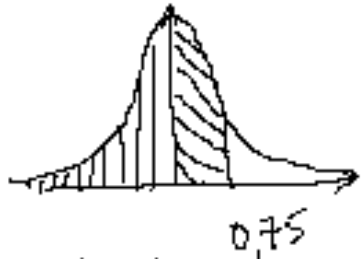
Envolvendo Normal.png



$2 \cdot (A(0,75) - 0,5)$



$A(0,75) - 0,5$



$A(0,75) = 0,7734$
da Tabela da Dist. Acumulada de Normal Padrão