Aproximação da Distribuição Binomial pela Normal

- Consideremos uma população em que a proporção de indivíduos portadores de uma certa característica é p.
- Colhida uma amostra casual simples de indivíduos (variáveis aleatórias independentes), podemos definir

$$X_i = \begin{cases} 1, & \text{se o indivíduo i possui a característica} \\ 0, & \text{caso contrário} \end{cases}$$

- $\Rightarrow X_i \sim \text{Bernoulli}(p);\ i = 1, 2, ..., n$
- Se as observações são independentes:

$$S_n = X_1 + ... + X_n \sim \text{binomial}(n, p)$$

Aproximação da Distribuição Binomial pela Normal

- $\hat{p} = \dfrac{S_n}{n}$ é uma média amostral.
- Utilizando a distribuição exata (n pequeno)

  $$P\left(\hat{p} = \frac{k}{n}\right) = P\left(\frac{S_n}{n} = \frac{k}{n}\right) = P\left(S_n = k\right) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k \left(1-p\right)^{n-k}$$

  $k = 0, 1, ..., n$
- Utilizando a aproximação para a Normal (n grande), temos que

  $$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Aproximação da Distribuição Binomial pela Normal

- **Exemplo:** Se p for a proporção de fumantes no estado de SP, $p = 0,2$ e tivermos coletado uma amostra casual simples de 500 indivíduos

$$X_i = \begin{cases} 1, & \text{se o indivíduo i é fumante} \\ 0, & \text{caso contrário} \end{cases}$$

  - $\hat{p} = \dfrac{\sum_{i=1}^{500} X_i}{500}$
  - $\hat{p} \sim N\left(0,2, \frac{0,2\times 0,8}{500}\right) = N\left(0,2, 0,00032\right)$.
  - $P\left(\hat{p} \leq 0,25\right) = P\left(Z \leq 2,795\right) = \Phi\left(2,795\right) = 0,9974$.

Aproximação da Distribuição Binomial pela Normal

- $\hat{p} = \dfrac{S_n}{n} \Rightarrow S_n = n\hat{p}.$
- Quando n é grande o suficiente $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$
- Qual a distribuição de S_n quando n é grande o suficiente?

Aproximação da Distribuição Binomial pela Normal

- Propriedade:
  - $X \sim N(a, b)$
  - $Y = \alpha X + \beta$
  - $\Rightarrow Y \sim N(\alpha a + \beta, \alpha^2 b)$
- Aplicação:
  - $S_n = X_1 + ... + X_n$
  - $\hat{p} = \frac{S_n}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$
  - $S_n = n\hat{p} \sim N\left(np, np(1-p)\right)$
- Portanto: binomial$(n, p) \approx N\left(np, np(1-p)\right)$ quando n é grande

Aproximação da Distribuição Binomial pela Normal

- **Exemplo:** $X \sim \text{binomial}(100; 0,4)$
  - $E(X) = 100 \times 0,4 = 40$
  - $Var(X) = 100 \times 0,4 \times 0,6 = 24$
  - $X \approx N(40, 24)$
  - $P(X \leq 50) = P\left(Z \leq \frac{50-40}{\sqrt{24}}\right) \approx \Phi\left(\frac{10}{\sqrt{24}}\right) = \Phi(2,04) \approx 0.9793$

Intervalo de Confiança como Estimativa de p

- $X \sim \text{binomial}(n, p)$
- n suficientemente grande, $X \sim N\left(np, np(1-p)\right)$
- $Z = \dfrac{X - np}{\sqrt{np(1-p)}} \approx N(0, 1)$
- $\gamma = 0,95$ é o grau de confiança

Intervalo de Confiança como Estimativa de p

$$
\begin{aligned}
0,95 &= P\left(-1,96 \leq Z \leq 1,96\right) \\
&= P\left(-1,96 \leq \frac{X - np}{\sqrt{np(1-p)}} \leq 1,96\right) \\
&= P\left(-1,96\sqrt{np(1-p)} \leq X - np \leq 1,96\sqrt{np(1-p)}\right) \\
&= P\left(\frac{-1,96\sqrt{np(1-p)}}{n} \leq \frac{X - np}{n} \leq \frac{1,96\sqrt{np(1-p)}}{n}\right) \\
&= P\left(\hat{p} - 1,96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1,96\sqrt{\frac{p(1-p)}{n}}\right)
\end{aligned}
$$

Intervalo de Confiança como Estimativa de p

- p é desconhecido
- $p(1-p) \leq \frac{1}{4}$
  - $\Rightarrow \sqrt{\dfrac{p(1-p)}{n}} \leq \sqrt{\dfrac{1}{4n}}$
  - $\Rightarrow -\sqrt{\dfrac{p(1-p)}{n}} \geq -\sqrt{\dfrac{1}{4n}}$
- $0,95 \approx P\left(\hat{p} - 1,96\sqrt{\frac{1}{4n}} \leq p \leq \hat{p} + 1,96\sqrt{\frac{1}{4n}}\right)$
- Caso geral:

  $\left[\hat{p} - z_\gamma\sqrt{\frac{1}{4n}}; \hat{p} + z_\gamma\sqrt{\frac{1}{4n}}\right]$ é um IC de $\gamma \times 100\%$ para p

Intervalo de Confiança como Estimativa de p

- **Exemplo:** Numa pesquisa de mercado, $n = 400$ pessoas foram entrevistadas sobre determinado produto, e 60% destas pessoas preferiam a marca A. Assim, $\hat{p} = 0,6$, logo, o IC com grau de confiança $\gamma = 0,95$ é dado por:

$$\left[0,6 - 1,96\frac{1}{\sqrt{1600}}; 0,6 + 1,96\frac{1}{\sqrt{1600}}\right] = [0,551; 0,649]$$

Intervalo de Confiança como Estimativa de p

- **Exemplo:** Suponha que em $n = 400$ provas, obtemos $k = 80$ sucesso. Vamos obter um intervalo de confiança para p, com $\gamma = 0,9$:
  - $\hat{p} = \dfrac{80}{400} = 0,2$
  - $z_{0,9} = 1,645$

$$\left[0,2 - 1,645\frac{1}{\sqrt{1600}}; 0,2 + 1,645\frac{1}{\sqrt{1600}}\right] = [0,159; 0,241]$$

  - Usando $\hat{p}$

$$\left[\hat{p} - z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0,167; 0,233]$$

Intervalo de Confiança como Estimativa de p

- O intervalo que utiliza $\hat{p}$ como estimativa tem menor amplitude do que o intervalo que utiliza $p(1-p) = \frac{1}{4}$
  - $[0,159; 0,2411]$: $0,2411 - 0,159 = 0,082$
  - $[0,167; 0,233]$: $0,233 - 0,167 = 0,066$
- Finalmente,os intervalos de confiança para p podem então ser de duas formas:

$$I_1 = \left[\hat{p} - z_\gamma\sqrt{\frac{1}{4n}}; \hat{p} + z_\gamma\sqrt{\frac{1}{4n}}\right]$$

$$I_2 = \left[\hat{p} - z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Intervalo de Confiança como Estimativa de p

- z_γ é tal que $\gamma = P\left(-z_\gamma \leq Z \leq z_\gamma\right)$; $Z \sim N(0,1)$
- Como determinar então, z_γ?

$$
\begin{aligned}
\gamma &= P\left(-z_\gamma \leq Z \leq z_\gamma\right) = P(Z \leq Z_\gamma) - P(Z \leq -Z_\gamma) \\
&= P(Z \leq Z_\gamma) - P(Z \geq Z_\gamma) = P(Z \leq Z_\gamma) - [1 - P(Z \leq -Z_\gamma)] \\
&= 2P(Z \leq Z_\gamma) - 1 = 2\Phi(Z_\gamma) - 1
\end{aligned}
$$

$$\Rightarrow \frac{\gamma + 1}{2} = \Phi(z_\gamma)$$

$$\Rightarrow \Phi^{-1}\left(\frac{\gamma + 1}{2}\right) = z_\gamma$$