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, tem-se que

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

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

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

Aproximação da Distribuição Binomial pela Normal

- Temos que $\hat{p} = \frac{S_n}{n}$ é a média amostral
- A distribuição exata (n pequeno) corresponde a

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

$k = 0, 1, ..., n$

- Utilizando a aproximação para a Normal (n grande), vem 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 vezes em que um determinado algoritmo não nos fornece a resposta de interesse em tempo hábil (com $p = 0.2$) e tivermos coletado uma amostra casual simples de 500 execuções, então

$$X_i = \begin{cases} 1, & \text{se o algoritmo não fornece a resposta de interesse em tempo h} \\ 0, & \text{caso contrário} \end{cases}$$

  - $\hat{p} = \frac{\sum_{i=1}^{500} X_i}{500}$
  - $\hat{p} \sim N\left(0.2, \frac{0.2 \times 0.8}{500}\right) = N\left(0.2, 0.00032\right)$
  - $P\left(\hat{p} \leq 0.25\right) = P\left(Z \leq 2.795\right) = \Phi\left(2.795\right) = 0.9974$

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:** $X \sim \text{binomial}(100, 0.4)$
  - $E(X) = 100 \times 0.4 = 40$
  - $Var(X) = 100 \times 0.4 \times 0.6 = 24$
  - $X \approx N(40, 24)$
  - $P(X \leq 50) = P\left(Z \leq \frac{50-40}{\sqrt{24}}\right) \approx \Phi\left(\frac{10}{\sqrt{24}}\right) = \Phi(2.04) \approx 0.9793.$
  - Usando a distribuição binomial (via R): pbinom(50,100,0.4) = 0,9832.

Intervalo de Confiança para p

- $X \sim Bin(n, p)$
- Para n suficientemente grande, então $X \sim N\left(np, np(1-p)\right)$
- $Z = \frac{X-np}{\sqrt{np(1-p)}} \approx N(0,1)$
- $\gamma = 0.95$ é o grau de confiança

Intervalo de Confiança para 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 para p

- p é desconhecido
- $p(1-p) \leq \frac{1}{4}$
  - $\Rightarrow \sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{1}{4n}}$
  - $\Rightarrow -\sqrt{\frac{p(1-p)}{n}} \geq -\sqrt{\frac{1}{4n}}$
- $0.95 \approx P\left(\hat{p} - 1.96\sqrt{\frac{1}{4n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{1}{4n}}\right)$
- 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 para p

- **Exemplo:** Voltando ao exemplo do algoritmo, suponha que em $n = 400$ execuções do algoritmo, em 60% dos casos ele forneceu a resposta desejada em tempo hábil. Assim $\hat{p} = 0.6$ e, portanto, um 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 para p

- **Exemplo:** Suponha que no exemplo anterior, em $n = 400$ execuções, obtivemos $k = 80$ sucessos. Vamos construir um intervalo de confiança para p, com $\gamma = 0.9$:
  - $\hat{p} = \frac{80}{400} = 0.2$ e $z_{0.9} = 1.645$.
  - Considerando (1/4n)

$$\left[0.2 - 1.645\frac{1}{\sqrt{1600}}; 0.2 + 1.645\frac{1}{\sqrt{1600}}\right] = [0.159; 0.241]$$

  - 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 para 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 para 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?

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

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

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

Intervalo de Confiança para a média populacional μ

- Temos $X_1, ..., X_n$ uma amostra aleatória de alguma distribuição com média e variância finitas (normal, exponencial, Poisson etc), ou seja $E(X) = \mu$ e $V(X) = \sigma^2$.
- Suponha que σ^2 é conhecido.
- Temos que $\overline{X} \sim N(\mu, \sigma^2/n)$ (mesmo se σ^2 for desconhecido).
- Portanto, $Z = \dfrac{\overline{X} - \mu}{\sqrt{\dfrac{\sigma^2}{n}}} \approx N(0,1)$, para n suficientemente grande.
- Se a distribuição dos dados for normal, o resultado acima é exato.

Intervalo de Confiança para μ, σ^2 conhecido

$$
\begin{aligned}
\gamma &= P\left(-z_\gamma \leq Z \leq z_\gamma\right) \\
&= P\left(-z_\gamma \leq \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \leq z_\gamma\right) \\
&= P\left(-z_\gamma\sqrt{\sigma^2/n} \leq \overline{X}-\mu \leq z_\gamma\sqrt{\sigma^2/n}\right) \\
&= P\left(\overline{X}-z_\gamma\sqrt{\sigma^2/n} \leq \mu \leq \overline{X}+z_\gamma\sqrt{\sigma^2/n}\right)
\end{aligned}
$$

Logo, $IC(\mu,\gamma) = \left[\overline{X}-z_\gamma\sqrt{\sigma^2/n}; \overline{X}+z_\gamma\sqrt{\sigma^2/n}\right]$

Intervalo de Confiança para a média populacional μ

- Suponha que σ^2 seja desconhecido.
- Além disso, $\dfrac{\overline{X}-\mu}{\sqrt{\dfrac{\widehat{\sigma}^2}{n}}} \approx t_{(n-1)}$, para n suficientemente grande, em que $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$.
- Se a distribuição dos dados for normal, o resultado acima é exato.
- Logo $IC(\mu,\gamma) = \left[\overline{X} - t_\gamma\sqrt{\widehat{\sigma}^2/n}; \overline{X} + t_\gamma\sqrt{\widehat{\sigma}^2/n}\right]$,analogamente aos desenvolvimentos feitos para o caso anterior.
- Em que $F^{-1}\left(\dfrac{\gamma+1}{2}, n-1\right) = t_\gamma$, em que $F(,n-1)$ é a fda de uma distribuição $t_{(n-1)}$.

Intervalo de Confiança para proporções

Exemplo

Suponha que $p = 30\%$ dos computadores de uma rede apresentam problemas de memória. Colhemos uma amostra aleatória simples de $n = 10$ computadores e calculamos $\hat{p}$ = proporção de computadores que apresentam problemas de memória. Qual a probabilidade de que $\hat{p}$ difira de p em menos de 0.01? E se $n = 50$?

Intervalo de Confiança para proporções

Temos que a probabilidade que desejamos encontrar é dada por

$$P\left(|\hat{p} - p| < 0.01\right) = P\left(-0.01 < \hat{p} - p < 0.01\right)$$

Onde p é o valor verdadeiro da proporção de computadores, e $\hat{p}$ a proporção observada na amostra. Sabemos que se n é grande, $\hat{p} - p$ pode ser aproximada por uma normal $N\left(0, p(1-p)/n\right)$. Como $p = 0.3$, temos que

$$\text{Var}\left(\hat{p} - p\right) = \frac{0.3 \cdot 0.7}{10} = 0.021$$

Intervalo de Confiança para proporções

Portanto, a probabilidade pedida é igual a

$$P\left(\frac{-0.01}{\sqrt{0.021}} < Z < \frac{0.01}{\sqrt{0.021}}\right) = P(-0.07 < Z < 0.07) = 0.056$$

Mas $n = 10$ é grande? Podemos comparar essa probabilidade com o resultado exato.

Não sabemos a distribuição de $\hat{p}$, mas o evento $\hat{p} = \alpha$ é igual ao evento $\sum X_i = n\alpha$, onde X_i são v.a. independentes e identicamente distribuidas Bernoulli(0.3). A soma é portanto binomial$(10, 0.3)$.

Intervalo de Confiança para proporções

O evento $\{|\hat{p} - p| < 0.01\}$ é igual ao evento $\{|\sum X_i - 10 \cdot 0.3| < 0.1\}$.

Como $\sum X_i$ assume somente valores inteiros, temos que

$$\left\{ \left| \sum_{i=1}^{10} X_i - 10 \cdot 0.3 \right| < 0.1 \right\} = \left\{ \sum_{i=1}^{10} X_i = 3 \right\}.$$

Portanto,

$$P\left(\left\{ \sum_{i=1}^{10} X_i = 3 \right\} \right) = \binom{10}{3} 0.3^3 0.7^7 = 0.267.$$

Temos uma probabilidade que é 5 vezes maior que a aproximação.

Intervalo de Confiança para proporções

Tome $n = 50$, agora. Podemos modificar rapidamente as contas da aproximação normal. A variância agora é 0.0042, e portanto a probabilidade aproximada é

$$P\left(\frac{-0.01}{\sqrt{0.0042}} < Z < \frac{0.01}{\sqrt{0.0042}}\right) = P(-0.154 < Z < 0.154) = 0.12239$$

A probabilidade exata agora é dada pelo evento $\{|\sum X_i - 50 \cdot 0.3| < 0.5\}$, ou simplesmente $\{\sum_{i=1}^{50} X_i = 15\}$.

Intervalo de Confiança para proporções

Observe agora que

$$P\left(\sum_{i=1}^{50} X_i = 15\right) = \binom{50}{15} 0.3^{15} 0.7^{50-15} = 0.12237$$

A diferença agora é muito menor e, é possível demonstrar, que à medida que $n \to \infty$ ela tende à 0. É preciso contudo ter em mente que a aproximação só é válida para grandes tamanhos de amostra, independentes e identicamente distribuidas.

Intervalo de Confiança para proporções

Exemplo

Uma amostra aleatória de 625 usuários revela que 70% deles estão satisfeitos com o provedor de internet que possuem. Construa um intervalo de confiança para p = proporção de usuários satisfeitos com coeficiente de confiança $\gamma = 90\%$.

Intervalo de Confiança para proporções

Temos que em nossa amostra aleatória $\hat{p} = 0.7$. Como $\hat{p} \sim N(p, p(1-p)/n)$, então o intervalo de confiança é dado por

$$\left(\hat{p} - z_\gamma \sqrt{\hat{p}(1-\hat{p})/n}\ ;\ \hat{p} + z_\gamma \sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Temos que para $\gamma = 0.90$, $z_\gamma = 1,64$ e portanto o intervalo de confiança para a proporção de interesse é dado por

$$\left(0.7 - 1.64\sqrt{0.7 \cdot 0.3/625}\ ;\ 0.7 + 1.64\sqrt{0.7 \cdot 0.3/625}\right)$$

$$(0.6699\ ;\ 0.7301)$$

Intervalo de Confiança para proporções

Exercício

Suponha que estejamos interessados em estimar a porcentagem de vezes em que um algoritmo estocástico retorna o valor abaixo de um certo limite de tempo. Se uma amostra de tamanho 300 (execuções desse algoritmo) forneceu 100 sucessos, determine:

(a) O intervalo de confiança de p, com c.c. de 95%; interprete o resultado.

(b) O tamanho da amostra para que o erro da estimativa não exceda 0.02 unidades com probabilidade de 95%; interprete o resultado.

Intervalo de Confiança para proporções

(a) O intervalo de confiança com 95% de confiabilidade é dado por:

$$\text{IC}(p; 0.95) = 0.333 \pm 1.96\sqrt{\frac{0.333 \cdot 0.667}{300}} = 0.333 \pm 0.053$$

, ou simplesmente $(0.280; 0.387)$.

Interpretação: Se pudéssemos construir um grande número de intervalos aleatórios para p, todos baseados em amostras do mesmo tamanho n, esperar-se-ia que 95% deles conteriam o parâmetro p.

Intervalo de Confiança para proporções

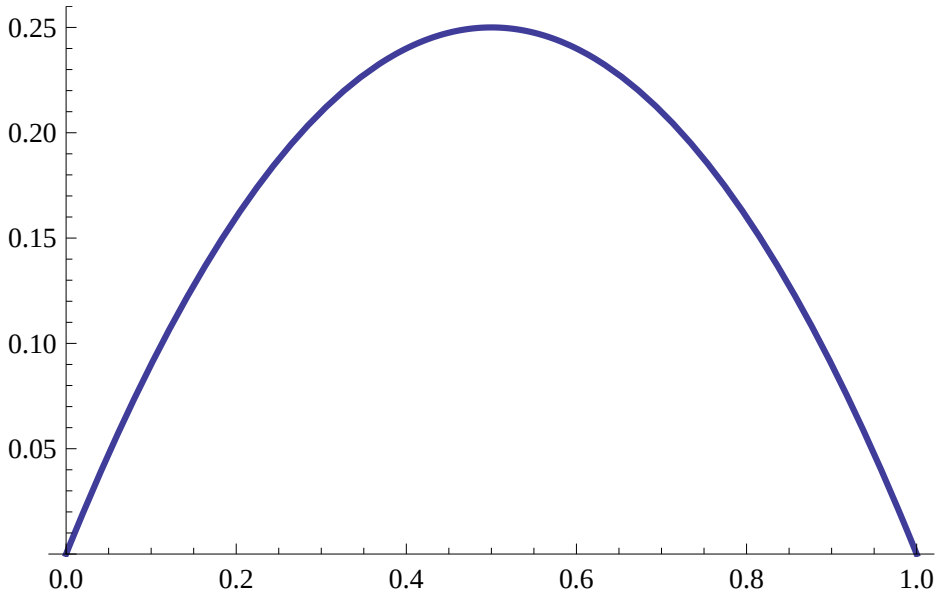
(b) Utilizando a estimativa da amostra observada ($\hat{p} = 0.333$), temos que n é dado por

$$n = \left(\frac{1.96}{0.02}\right)^2 \times 0.333 \times 0.667 \cong 2134.$$

Contudo, frequentemente devemos determinar o tamanho da amostra antes de realizar qualquer experimento, isto é, sem nenhuma informação prévia de p. Se esse for o caso, devemos considerar que a variância da amostra é a maior possível.

Intervalo de Confiança para proporções

(b) Se considerarmos a variância como função de p, obtemos o seguinte gráfico:



Note que a variância é máxima quando $p = 1/2$.

Intervalo de Confiança para proporções

(b) Utilizando o valor máximo de $p(1-p)$, isto é, $1/4$, obtemos

$$n = \left(\frac{1.96}{0.02}\right)^2 \times \frac{1}{4} \cong 2401$$

Interpretação: Utilizando o tamanho amostral encontrado, teremos uma probabilidade de 95% de que a proporção amostral não difira do verdadeiro valor de p em menos que 2%.

Obter amostras pequenas para examinar p, e então determinar o tamanho amostral sem utilizar o "pior caso", ´é no que consiste a idéia de *amostras piloto*.

Intervalo de Confiança para a média

a) Uma empresa de tecnologia da informação objetiva saber qual o tempo médio de varredura de um anti-vírus (AV) . Para isso coletou uma amostra de 50 tempos (em minutos) em que tal AV foi utilizado, em diferentes computadores, resultando em $\overline{x} = 14,5$. Supondo $\sigma^2 = 100$, construa um $IC(\mu, 0,99)$.

b) Repita o item [a)] considerando σ^2 desconhecido, e $\widetilde{\sigma}^2 = 110,87$.

Intervalo de Confiança para a média

a) $z_\gamma = 2,57,\ IC(\mu, 0,99) =$

$[14,5 - 2,57 \times 10/\sqrt{50}; 14,5 + 2,57 \times 10/\sqrt{50}] = [10,87; 18,13].$

b) $t_\gamma = 2,704,\ IC(\mu, 0,99) = [14,5 - 2,704 \times \sqrt{110,87}/\sqrt{50}; 14,5 +$

$2,704 \times \sqrt{110,87}/\sqrt{50}] = [10,47; 18,53].$

Intervalo de Confiança para diferenças entre médias

Exemplo

Estão sendo estudados dois processadores, em relação a velocidade de processamento em determinada unidade. No processador A, a velocidade X de duração segue uma distribuição $N(\mu_A, 100)$, e no processador B a velocidade Y obedece à distribuição $N(\mu_B, 100)$. Consideram-se duas amostras (testes feitos em diferentes computadores com configurações semelhantes): a de A, com 16 computadores, apresentou tempo médio de duração igual a 50, e a de B, com 25 computadores, duração média igual a 60.

(a) Construa um IC para μ_A e μ_B, separadamente.

(a) Para o caso geral, o intervalo de confiança para μ, para variância conhecida, com coeficiente de confiabilidade γ, é dado por

$$\left(\bar{X} - z_\gamma \sqrt{\sigma^2/n}\ ;\ \bar{X} + z_\gamma \sqrt{\sigma^2/n}\right)$$

Note que $\sigma_A = \sigma_B$. Para o coeficiente de confiança $\gamma = 0.95$, por exemplo, temos $z_\gamma = 1.96$, e os intervalos de confiança serão, respectivamente:

$$\text{IC}(\mu_A) = \left(50 - 1.96\sqrt{100/16}\ ;\ 50 + 1.96\sqrt{100/16}\right)$$

$$\text{IC}(\mu_B) = \left(60 - 1.96\sqrt{100/25}\ ;\ 60 + 1.96\sqrt{100/25}\right)$$

Intervalo de Confiança

(a) (cont.) Fazendo as contas, obtemos que

$$\text{IC}(\mu_A) = (45.1 \ ; \ 54.9)$$

$$\text{IC}(\mu_B) = (56.08 \ ; \ 63.92)$$

Observe que os intervalos não se interceptam; temos evidência de que os tempos de processamentos são diferentes, a 95% de confiança, com vantagem para o processador A.

Intervalo de Confiança

(b) Repita o procedimento considerando as variâncias desconhecidas, tais que $\widetilde{\sigma}_A^2 = 90,34$ e $\widetilde{\sigma}_B^2 = 85,39$.

Nesse caso, para o processador A (n-1 = 15) $t_\gamma = 2,131$ e para o processador B (n-1 = 24) $t_\gamma = 2,064$. Portanto

$$\text{IC}(\mu_A) = \left(50 - 2,131\sqrt{90,34/16}\ ;\ 50 + 2,131\sqrt{90,34/16}\right)$$

$$= (44,94; 55,06)$$

$$\text{IC}(\mu_B) = \left(60 - 2,064\sqrt{85,39/25}\ ;\ 60 + 2,064\sqrt{85,39/25}\right)$$

$$= (56,19; 63,81)$$