Comparação de duas proporções

Como vimos para a média, muito freqüentemente, podemos estar interessados na *comparação de duas proporções de duas populações independentes*.

- Hipóteses estatísticas:

$$\boldsymbol{H}: p_1 = p_2$$
$$\boldsymbol{A}: p_1 \neq p_2$$

$\rightarrow$ extraímos uma uma a.a. de tamanho n_1 de uma população com proporção p_1; se observamos x_1 sucessos na amostra, então

$$\hat{p}_1 = \frac{x_1}{n_1}$$

$\rightarrow$ Analogamente, selecionamos uma amostra de tamanho n_2 da população com proporção p_2 e se observamos x_2 sucessos, então

$$\hat{p}_2 = \frac{x_2}{n_2}$$

- Estimador de $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \quad \Rightarrow \quad \begin{cases} E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \\ Var(\hat{p}_1 - \hat{p}_2) = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2} \end{cases}$$

A fim de verificar se a diferença observada nas proporções das amostras, $\hat{p}_1 - \hat{p}_2$, é muito grande para ser atribuída apenas ao acaso, calculamos, então, a probabilidade de se obter um par de proporções tão ou mais discrepante que aquele observado, dado que a hipótese nula *H* seja verdadeira (nível descritivo ***P***).

Para um nível de significância α especificado:

P $\leq \alpha \Rightarrow$ rejeitamos *H* e concluímos que as duas proporções das populações são diferentes.

P $> \alpha \Rightarrow$ não rejeitamos *H.*

Se a hipótese nula é verdadeira, temos que $p_1 = p_2 = p$, os dados de ambas as amostras podem ser combinados para estimar esse parâmetro comum, por

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

A quantidade $\hat{p}$ é uma média ponderada das duas proporções das amostras, $\hat{p}_1$ e $\hat{p}_2$.

Sob a hipótese nula, o estimador do *erro padrão da diferença* $\hat{p}_1 - \hat{p}_2$ é dado por:

$$\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

• estatística do teste:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Se n_1 e n_2 são suficientemente grandes, essa estatística, sob *H*, tem uma distribuição normal com média 0 e desvio padrão 1.

Exemplo: Em um estudo que investiga mortalidade entre vítimas pediátricas de acidentes de veículos a motor, a informação com relação à efetividade dos cintos de segurança foi coletada em um período de 18 meses. Duas a.a. foram selecionadas, uma da população de crianças que estavam usando cintos de segurança no momento do acidente e outra da população que não estavam.

Queremos testar $H: p_1 = p_2$ *versus* $A: p_1 \neq p_2$,

ou seja, a hipótese nula de que as proporções de crianças que morreram como resultado do acidente são idênticas nas duas populações.

Amostra 1: n_1=123 crianças que estavam usando cinto de segurança no momento do acidente $\Rightarrow$ 3 morreram $\Rightarrow \hat{p}_1 = 0{,}024$

Amostra 2: n_2=290 crianças que não estavam usando cinto de segurança $\Rightarrow$ 13 morreram $\Rightarrow \hat{p}_2 = 0{,}045$

$\rightarrow$ Essa discrepância nas proporções das amostras – uma diferença de 0,021 ou 2,1% – é muito grande para ser atribuída ao acaso?

- teste bilateral ao nível de significância de 0,05

Se as proporções das populações, p_1 e p_2, são de fato iguais, seu valor *p* comum é estimado por 0,039.

- Valor da estatística do teste:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$ = -1,01, substituindo os valores.

- Nível descritivo:

$P = 2 \times P(Z \leq -1{,}01) = 0{,}312 \Rightarrow$ não rejeitamos *H.*

$\Rightarrow$ As amostras coletadas nesse estudo particular não fornecem evidência de que as proporções de crianças que morrem diferem entre aquelas que estavam usando cinto de segurança e aquelas que não estavam.

$\hat{p}_1 - \hat{p}_2$ $\Rightarrow$ fornece uma *estimativa por ponto* para a verdadeira diferença $p_1 - p_2$ das proporções populacionais.

Um *intervalo de confiança de 95%* para a diferença $p_1 - p_2$, usando a aproximação normal, é

$$\left(\hat{p}_1 - \hat{p}_2 \mp 1{,}96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

Note que o erro padrão da diferença das proporções amostrais não é o mesmo que aquele usado no teste; no teste de hipóteses, o erro padrão empregado, foi baseado na suposição de que a hipótese nula era verdadeira; essa suposição não é necessária no cálculo de um intervalo de confiança.

No exemplo, como $\hat{p}_1 = 0{,}024$ e $\hat{p}_2 = 0{,}045$, um intervalo de confiança aproximado de 95% para $p_1 - p_2$ é

$$\left((0{,}024 - 0{,}045) \mp 1{,}96\sqrt{\frac{0{,}024(1-0{,}024)}{123} + \frac{0{,}045(1-0{,}045)}{290}}\right)$$

$$(-0{,}057\ ;\ 0{,}015\ )$$

(Note que, como esperado, o intervalo contém o valor zero)