4.6.2. Teste da diferença entre proporções

Se quisermos comparar as *proporções de sucessos* de duas populações, iremos tirar uma amostra (grande) de cada, e fazer inferência a partir da diferença entre as proporções de sucesso encontradas nestas amostras.

4.6.2.1. Distribuição amostral da diferença entre proporções

Como no caso da diferença entre médias (seção 4.6.1), aqui também é fácil ver que a diferença D entre as proporções P_1 e P_2 de sucessos de duas amostras grandes também terá distribuição normal: a proporção de sucessos em uma amostra tem distribuição que tende para a normal, quando cresce o tamanho da amostra (seção 3.4.4.5), e a diferença entre duas variáveis normais também é uma variável normal (seção 3.4.4.3).

Os parâmetros da distribuição das diferenças D também são fáceis de deduzir, dadas as propriedades do valor esperado e da variância da diferença entre duas variáveis aleatórias (seção 3.3.1.5):

- o valor esperado da diferença $D = P_1 - P_2$ é a diferença entre os valores esperados das variáveis P_1 e P_2:

$$D = P_1 - P_2$$
$$E(D) = E(P_1) - E(P_2) = \pi_1 - \pi_2$$

- a variância da diferença D é a soma das variâncias das variáveis P_1 e P_2:

$$V(P_1) = \frac{\pi_1(1-\pi_1)}{n_1}$$
$$V(P_2) = \frac{\pi_2(1-\pi_2)}{n_2}$$
$$V(D) = V(P_1) + V(P_2)$$
$$V(D) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Estes resultados podem ser enunciados como no Teorema **1**.

Teorema 1. *Distribuição amostral da diferença de proporções D*
Se duas populações têm proporções de sucessos representadas por π_1 e π_2, e retiramos delas amostras aleatórias de tamanhos n_1 e n_2 a diferença entre as proporções de sucessos P_1 e P_2 destas amostras,

$$D = P_1 - P_2$$

terá distribuição que tende para a normal:

$$D \rightarrow N(\mu_D, \sigma_D^2) \text{ quando } n_1 \rightarrow \infty \text{ e } n_2 \rightarrow \infty$$

com parâmetros:

$$\mu_D = \pi_1 - \pi_2 \quad \text{e} \quad \sigma_D = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

A variável D pode ser transformada em uma variável de teste padronizada $Z \sim N(0,1)$, por meio de:

$$Z = \frac{D - \mu_D}{\sigma_D} = \frac{D - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}}$$

4.6.2.2. Teste de hipótese sobre diferença de proporções

Na maioria dos testes feitos para comparar proporções, a hipótese nula é a de que as duas populações têm a mesma proporção de sucessos, e portanto a diferença entre elas é nula:

$$H_0 : \ \pi_1 - \pi_2 = 0$$

Como as proporções de sucessos na população (π_1 e π_2) são desconhecidos, devem ser estimadas a partir das proporções observadas nas amostras (P_1 e P_2). O teste será portanto baseado numa variável de teste padronizada Z, calculada por:

$$Z = \frac{D - \mu_D}{\sigma_D} = \frac{D - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}} \cong \frac{D}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_2)}{n_2}}} \qquad (1)$$

4.6.2.3. Exemplos

(i) Exemplo 1:
Proporção de nascimentos de crianças com baixo peso, de acordo com a raça da mãe

O fato de nascer com baixo peso (peso ≤ 2.500 g) é considerado um dos fatores de risco para a saúde de uma criança recém-nascida. Para investigar se a probabilidade de uma criança nascer com baixo peso está associada à raça da mãe, nos EUA, foi analisada uma amostra de amostra de 189 nascimentos (já usada no exemplo da seção 4.6.1.3). As mães foram classificadas pela raça em duas categorias, *brancas* e *não-brancas*. A Tab. **1** mostra a porcentagem de mães em cada raça.

Tabela 1. Raça das mães

	f	fr(%)
branca	96	50,8
não-branca	93	49,2
total	189	100

A Tab. **2** mostra os pesos na nascer das crianças, classificados em duas classes, *normal* e *baixo*. Nestas amostras, as mães brancas tiveram uma proporção menor de crianças nascidas com baixo peso (0,24) do que as mães não-brancas (0,387). Faremos um teste de diferença de proporções para verificar se esta diferença observada nas amostras é significativa (isto é, se pode ser generalizada para toda a população).

Tabela 2. Peso ao nascer da criança (normal / baixo) pela raça da mãe

	raça da mãe				Total
	(1) brancas		(2) não-brancas		
peso ao nascer	f	fr	f	fr%	
normal	73	0,760	57	0,613	130
baixo	23	0,240	36	0,387	59
Total	96	1,000	93	1,000	189

Iremos testar a hipótese de que as mães brancas têm menor proporção de crianças de baixo peso. As hipótese nula e alternativa serão definidas como:

$$H_0: \quad \pi_1 - \pi_2 \geq 0$$

$$H_1: \quad \pi_1 - \pi_2 < 0$$

onde

π_1 : proporção na população de mães brancas que têm crianças com baixo peso

π_2 : proporção na população de mães não-brancas que têm crianças com baixo peso

O teste será unilateral, e usaremos α=0,05, o que leva a um valor crítico Z_c = -1,96. As estatísticas observadas nas amostras são:

P_1 : proporção na amostra de mães brancas que tiveram crianças com baixo peso

P_2 : proporção na amostra de mães não-brancas que tiveram crianças com baixo peso

n_1 : tamanho da amostra 1 (mães brancas)

n_2 : tamanho da amostra 2 (mães não-brancas)

Cujos valores são

$P_1 = 0{,}240$ $\quad n_1 = 96$ $\quad P_2 = 0{,}387$ $\quad n_2 = 93$

Para calcular o valor da estatística de teste Z, usamos a eq. (**1**)

$$Z \cong \frac{D}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_2)}{n_2}}} = \frac{0{,}24 - 0{,}387}{\sqrt{\dfrac{0{,}24(1-0{,}24)}{96} + \dfrac{0{,}387(1-0{,}387)}{93}}} = -2{,}20$$

O valor de *Z* encontrado está abaixo do valor crítico, $Z < Z_c$, e portanto dentro da área de rejeição; concluímos então pela rejeição da H_0 e aceitação da H_1: as mães brancas nesta população têm uma proporção de crianças de baixo *menor* do que as mães não-brancas. Este resultado não é muito significativo, porém; calculando com o R, encontramos um valor-p = 0,0139, que é significativo para α=0,05, mas não para α=0,01 (se α=0,01, o valor crítico será Z_c = -2,33, e *Z* não irá alcançar a área de rejeição).