

Elementarna statystyka

Dwie próby: porównanie dwóch proporcji

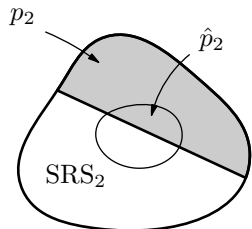
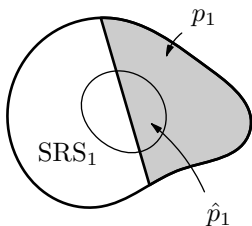
(Two-sample problem: comparing two proportions)

Alexander Bendikov

Uniwersytet Wrocławski

25 maja 2016

Dwie próby: porównanie dwóch proporcji



W zagadnieniu dwóch prób chcemy porównać dwie populacje, na przykład reakcję na terapię w dwóch populacjach, na podstawie dwóch niezależnie pobranych prób:

$$SRS_1 : p_1, n_1, \hat{p}_1;$$

$$SRS_2 : p_2, n_2, \hat{p}_2,$$

gdzie będziemy testować hipotezę:

$$\begin{cases} H_0 : p_1 = p_2, \\ H_a : p_1 \neq p_2 \quad (<, >). \end{cases}$$

Przykład: Organizacja ERF (Educational Research Foundation) postanowiła zbadać długofalowe skutki uczęszczania dzieci do przedszkola. Obserwowano w długim okresie 2 grupy dzieci, poczynając od wczesnego dzieciństwa. Pierwsza, kontrolna, grupa składała się z $n_1 = 61$ dzieci, wybranych losowo. Druga grupa, o liczności $n_2 = 62$ składała się z dzieci wybranych losowo z tego samego środowiska, ale spośród dzieci uczęszczających do przedszkola w wieku 3–4 lat. Zmienną badaną było zapotrzebowanie badanych dzieci na pomoc społeczną (Social Services) w ich późniejszym, dorosłym życiu (zmienna SS).

Przykład: Organizacja ERF (Educational Research Foundation) postanowiła zbadać długofalowe skutki uczęszczania dzieci do przedszkola. Obserwowano w długim okresie 2 grupy dzieci, poczynając od wczesnego dzieciństwa. Pierwsza, kontrolna, grupa składała się z $n_1 = 61$ dzieci, wybranych losowo. Druga grupa, o liczności $n_2 = 62$ składała się z dzieci wybranych losowo z tego samego środowiska, ale spośród dzieci uczęszczających do przedszkola w wieku 3–4 lat. Zmienną badaną było zapotrzebowanie badanych dzieci na pomoc społeczną (Social Services) w ich późniejszym, dorosłym życiu (zmienna SS). Okazało się, że w ciągu ostatnich 10 lat w grupie osób uczęszczających w dzieciństwie do przedszkola 38 wymagało pomocy społecznej, a w grupie kontrolnej 49.

Przykład: Organizacja ERF (Educational Research Foundation) postanowiła zbadać długofalowe skutki uczęszczania dzieci do przedszkola. Obserwowano w długim okresie 2 grupy dzieci, poczynwszy od wczesnego dzieciństwa. Pierwsza, kontrolna, grupa składała się z $n_1 = 61$ dzieci, wybranych losowo. Druga grupa, o liczności $n_2 = 62$ składała się z dzieci wybranych losowo z tego samego środowiska, ale spośród dzieci uczęszczających do przedszkola w wieku 3–4 lat. Zmienną badaną było zapotrzebowanie badanych dzieci na pomoc społeczną (Social Services) w ich późniejszym, dorosłym życiu (zmienna SS). Okazało się, że w ciągu ostatnich 10 lat w grupie osób uczęszczających w dzieciństwie do przedszkola 38 wymagało pomocy społecznej, a w grupie kontrolnej 49.

Grupa	n	SS	$\hat{p}$
Kontrolna	61	49	0,803
Przedszkolna	62	38	0,613

$$\hat{p}_1 = \frac{49}{61} = 0,803 \quad \longrightarrow \quad \text{grupa kontrolna}$$

$$\hat{p}_2 = \frac{38}{62} = 0,613 \quad \longrightarrow \quad \text{grupa przedszkolna}$$

Na podstawie danych wykonamy 2 badania statystyczne:

$$\hat{p}_1 = \frac{49}{61} = 0,803 \quad \longrightarrow \quad \text{grupa kontrolna}$$

$$\hat{p}_2 = \frac{38}{62} = 0,613 \quad \longrightarrow \quad \text{grupa przedszkolna}$$

Na podstawie danych wykonamy 2 badania statystyczne:

1. Aby oszacować jak duży wpływ ma wychowanie przedszkolne na późniejsze problemy utworzymy przedział ufności dla zmiennej $p_1 - p_2$ będącej różnicą proporcji w dwóch badanych populacjach

$$\hat{p}_1 = \frac{49}{61} = 0,803 \quad \longrightarrow \quad \text{grupa kontrolna}$$

$$\hat{p}_2 = \frac{38}{62} = 0,613 \quad \longrightarrow \quad \text{grupa przedszkolna}$$

Na podstawie danych wykonamy 2 badania statystyczne:

1. Aby oszacować jak duży wpływ ma wychowanie przedszkolne na późniejsze problemy utworzymy przedział ufności dla zmiennej $p_1 - p_2$ będącej różnicą proporcji w dwóch badanych populacjach
2. Aby sprawdzić, czy dane dostarczają istotnego dowodu na to, że uczęszczanie do przedszkola w dzieciństwie zmniejsza w przyszłości zapotrzebowanie na pomoc społeczną będziemy testowali hipotezy:

$$H_0 : \quad p_1 = p_2, \quad H_a : \quad p_1 > p_2.$$

W obu przypadkach zaczynamy od $\hat{p}_1 - \hat{p}_2 = 0,190$. Przypomnijmy, że pojedyncze próbki X_1, X_2 w obu grupach mają rozkład bliski rozkładowi dwumianowemu z $E(X_i) = p_i$, $Var(X_i) = p_i(1 - p_i)$. Ponieważ zmienne są niezależne, więc mamy następujące statystyczne parametry próby:

- $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$,
- $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$,
- Z-statystyka:

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

W obu przypadkach zaczynamy od $\hat{p}_1 - \hat{p}_2 = 0,190$. Przypomnijmy, że pojedyncze próbki X_1, X_2 w obu grupach mają rozkład bliski rozkładowi dwumianowemu z $E(X_i) = p_i$, $Var(X_i) = p_i(1 - p_i)$. Ponieważ zmienne są niezależne, więc mamy następujące statystyczne parametry próby:

- $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$,
- $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$,
- Z-statystyka:

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

Główne założenia:

1. Dwie badane populacje są niezależne,
2. $n_i \hat{p}_i, n_i(1 - \hat{p}_i) \geq 10$. To uzasadnia użycie CLT, i możemy przyjąć, że $\hat{p}_i$ mają rozkład normalny. W takim razie różnica $\hat{p}_1 - \hat{p}_2$ też ma rozkład normalny, a więc Z-statystyka jest w przybliżeniu normalną, standardową zmienną losową $N(0, 1)$.

Przedział ufności:

$$(\hat{p}_1 - \hat{p}_2) \pm z_* SE, \quad \text{gdzie } SE \simeq \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}$$

Nieznane parametry p_i zastąpiliśmy przez parametry próby $\hat{p}_i$. Wybierzmy poziom ufności $C = 95\%$

- (1) $n_1\hat{p}_1 = 49$, $n_1(1 - \hat{p}_1) = 12$; $n_2\hat{p}_2 = 38$, $n_2(1 - \hat{p}_2) = 14$,
- (2) $SE \approx (0,803 \cdot 0,197/61 + 0,613 \cdot 0,387/62)^{1/2} = 0,0801$,
- (3) Przedział ufności: $(0,803 - 0,613) \pm 1,960 \cdot 0,0801 = 0,190 \pm 0,157$:

$$0,033 < p_1 - p_2 < 0,347.$$

Testowanie hipotezy:

$$H_0 : p_1 = p_2, \quad H_a : p_1 > p_2.$$

(1) Załóżmy H_0 , mamy

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE} = \frac{0,190}{\sqrt{\frac{0,803 \cdot 0,197}{61} + \frac{0,613 \cdot 0,387}{62}}} = \frac{0,190}{0,0801} = 2,372.$$

Ponownie, nieznanne parametry p_i zastąpiliśmy przez parametry próby $\hat{p}_i$.

Testowanie hipotezy:

$$H_0 : p_1 = p_2, \quad H_a : p_1 > p_2.$$

(1) Załóżmy H_0 , mamy

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE} = \frac{0,190}{\sqrt{\frac{0,803 \cdot 0,197}{61} + \frac{0,613 \cdot 0,387}{62}}} = \frac{0,190}{0,0801} = 2,372.$$

Ponownie, nieznanne parametry p_i zastąpiliśmy przez parametry próby $\hat{p}_i$.

(2) p -wartość $= P(z > 2,372) = 0,0089 \simeq 0,009 = 0,9\% < \alpha$.

Testowanie hipotezy:

$$H_0 : p_1 = p_2, \quad H_a : p_1 > p_2.$$

(1) Załóżmy H_0 , mamy

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE} = \frac{0,190}{\sqrt{\frac{0,803 \cdot 0,197}{61} + \frac{0,613 \cdot 0,387}{62}}} = \frac{0,190}{0,0801} = 2,372.$$

Ponownie, nieznane parametry p_i zastąpiliśmy przez parametry próby $\hat{p}_i$.

- (2) p -wartość $= P(z > 2,372) = 0,0089 \simeq 0,009 = 0,9\% < \alpha$.
- (3) Widzimy, że przy jakimkolwiek poziomie istotności $\alpha \geq 0,01$ należy odrzucić H_0 na rzecz H_a , $p_1 > p_2$, czyli dane potwierdzają, że uczestnictwo w programie przedszkolnym w dzieciństwie zmniejsza późniejsze zapotrzebowanie na pomoc społeczną.

Test dla próby połączonej: Często przy testowaniu hipotezy $p_1 = p_2$ stosuje się metodę „próby połączonej”. Skoro testujemy przy założeniu takiej samej proporcji w obu populacjach $p = p_1 = p_2$, to możemy to wykorzystać w obliczaniu przybliżenia:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}.$$

Z-statystyka przyjmuje postać:

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

Test dla próby połączonej: Często przy testowaniu hipotezy $p_1 = p_2$ stosuje się metodę „próby połączonej”. Skoro testujemy przy założeniu takiej samej proporcji w obu populacjach $p = p_1 = p_2$, to możemy to wykorzystać w obliczaniu przybliżenia:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}.$$

Z-statystyka przyjmuje postać:

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

W naszym przypadku:

- $\hat{p} = \frac{49+38}{61+62} = 0,707,$
- $z = \frac{0,190}{\sqrt{0,707 \cdot 0,293 \left(\frac{1}{61} + \frac{1}{62}\right)}} = 2,31$ (porównajmy to z poprzednio
otrzymaną wartością 2,372. Wartość 2,31 jest dokładniejsza.
- $p = 0,0104$. Przy poziomie istotności, powiedzmy $\alpha = 0,05$ odrzucamy hipotezę H_0 na rzecz H_a : $p_1 > p_2$.

Przykład: Niektóre badania sugerują, że aspiryna może redukować ryzyko zawałów i udarów. W dużym badaniu utworzono dwie grupy obserwowanych przypadków. Pacjentom jednej grupy podawano regularnie aspirynę, a pacjentom drugiej grupy placebo. Otrzymano następujące dane:

Grupa terapeutyczna	Aspiryna	Placebo
Zawały	129	213
Udary	119	98
Liczność grup	11037	11034

Przykład: Niektóre badania sugerują, że aspiryna może redukować ryzyko zawałów i udarów. W dużym badaniu utworzono dwie grupy obserwowanych przypadków. Pacjentom jednej grupy podawano regularnie aspirynę, a pacjentom drugiej grupy placebo. Otrzymano następujące dane:

Grupa terapeutyczna	Aspiryna	Placebo
Zawały	129	213
Udary	119	98
Liczność grup	11037	11034

Czy te dane dostarczają statystycznie istotnego dowodu na sugestię, że zażywanie aspiryny redukuje ryzyko zawału lub udaru?

Przykład: Niektóre badania sugerują, że aspiryna może redukować ryzyko zawałów i udarów. W dużym badaniu utworzono dwie grupy obserwowanych przypadków. Pacjentom jednej grupy podawano regularnie aspirynę, a pacjentom drugiej grupy placebo. Otrzymano następujące dane:

Grupa terapeutyczna	Aspiryna	Placebo
Zawały	129	213
Udary	119	98
Liczność grup	11037	11034

Czy te dane dostarczają statystycznie istotnego dowodu na sugestię, że zażywanie aspiryny redukuje ryzyko zawału lub udaru?

1) Zawały: mamy:

$$\begin{aligned}n_1 &= 11037, & n_2 &= 11034, \\ \hat{p}_1 &= 129/11037 = 0,01169, \\ n_1 \hat{p}_1 &= 129, & n_1(1 - \hat{p}_1) &= 10908, \\ \hat{p}_2 &= 213/11034 = 0,01930, \\ n_2 \hat{p}_2 &= 213, & n_2(1 - \hat{p}_2) &= 10821.\end{aligned}$$

Wszystkie liczby $n_i\hat{p}_i$, $n_i(1 - \hat{p}_i)$ są większe niż 10, więc możemy zastosować Z-test.

(a) Hipotezy: $H_0 : p_1 = p_2$, $H_a : p_1 \neq p_2$.

(b) Proporcja w próbie połączonej:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{129 + 213}{11037 + 11034} = 0,01550.$$

Wszystkie liczby $n_i\hat{p}_i$, $n_i(1 - \hat{p}_i)$ są większe niż 10, więc możemy zastosować Z-test.

(a) Hipotezy: $H_0 : p_1 = p_2$, $H_a : p_1 \neq p_2$.

(b) Proporcja w próbie połączonej:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{129 + 213}{11037 + 11034} = 0,01550.$$

(c) Z-test:

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \\ &= \frac{0,01169 - 0,01930}{\sqrt{0,0155(1 - 0,0155)\left(\frac{1}{11037} + \frac{1}{11034}\right)}} \simeq -458. \end{aligned}$$

(d) p -wartość: mniej niż 0,00001 (!).

Wszystkie liczby $n_i\hat{p}_i$, $n_i(1 - \hat{p}_i)$ są większe niż 10, więc możemy zastosować Z-test.

(a) Hipotezy: $H_0 : p_1 = p_2$, $H_a : p_1 \neq p_2$.

(b) Proporcja w próbie połączonej:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{129 + 213}{11037 + 11034} = 0,01550.$$

(c) Z-test:

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \\ &= \frac{0,01169 - 0,01930}{\sqrt{0,0155(1 - 0,0155)\left(\frac{1}{11037} + \frac{1}{11034}\right)}} \simeq -458. \end{aligned}$$

(d) p -wartość: mniej niż 0,00001 (!).

(e) Wniosek: Dane dostarczają bardzo mocnego dowodu (na poziomie istotności $\alpha \geq 10^{-5}$) że grupa leczona aspiryną i grupa „leczona” placebem (?) miały inne ryzyko zawału. (Grupa leczona aspiryną miała

2) Udary: podobnie jak dla zawałów, obliczamy

$$n_1 = 11037, \quad n_2 = 11034,$$

$$\hat{p}_1 = 119/11037 = 0,01078,$$

$$n_1\hat{p}_1 = 119, \quad n_1(1 - \hat{p}_1) = 10918,$$

$$\hat{p}_2 = 98/11034 = 0,00888,$$

$$n_2\hat{p}_2 = 98, \quad n_2(1 - \hat{p}_2) = 10936.$$

Tak jak poprzednio, wszystkie liczby $n_i\hat{p}_i$, $n_i(1 - \hat{p}_i)$ są większe niż 10, więc możemy zastosować Z-test.

(a) Hipotezy: $H_0 : p_1 = p_2$, $H_a : p_1 \neq p_2$.

(b) Proporcja w próbie połączonej:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{119 + 98}{11037 + 11034} = 0,009832.$$

2) Udary: podobnie jak dla zawałów, obliczamy

$$\begin{aligned}n_1 &= 11037, & n_2 &= 11034, \\ \hat{p}_1 &= 119/11037 = 0,01078, \\ n_1\hat{p}_1 &= 119, & n_1(1 - \hat{p}_1) &= 10918, \\ \hat{p}_2 &= 98/11034 = 0,00888, \\ n_2\hat{p}_2 &= 98, & n_2(1 - \hat{p}_2) &= 10936.\end{aligned}$$

Tak jak poprzednio, wszystkie liczby $n_i\hat{p}_i$, $n_i(1 - \hat{p}_i)$ są większe niż 10, więc możemy zastosować Z-test.

(a) Hipotezy: $H_0 : p_1 = p_2$, $H_a : p_1 \neq p_2$.

(b) Proporcja w próbie połączonej:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{119 + 98}{11037 + 11034} = 0,009832.$$

(c) Z-test:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,01078 - 0,00888}{\sqrt{0,0098 \cdot 0,99\left(\frac{1}{11037} + \frac{1}{11034}\right)}} \simeq 1,43.$$

- (4) p -wartość: $p = 2P(z > 1,43) = 0,1528$.
- (5) Wniosek: Dane nie dostarczają dowodu że grupa leczona aspiryną i grupa „leczona” placebem (?) miały inne ryzyko udaru.