Porównanie dwu populacji

Porównanie dwóch rozkładów normalnych

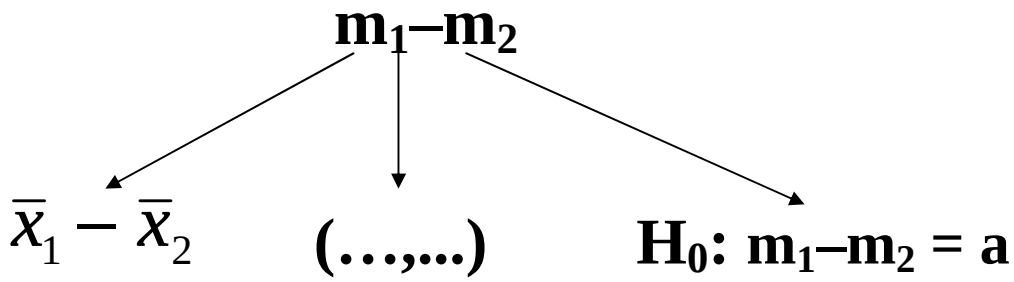
Założenia:

1. $X_1 \sim N(m_1, s_1^2)$, $X_2 \sim N(m_2, s_2^2)$, $\quad s_1^2 = s_2^2$

2. parametry rozkładów nie są znane

2. *X_1, X_2 są niezależne.*

Ocena różnicy między średnimi $m_1 - m_2$



$$\{(\bar{x}_1 - \bar{x}_2) - t(a, n_1 + n_2 - 2)s_r ; (\bar{x}_1 - \bar{x}_2) + t(a, n_1 + n_2 - 2)s_r\}$$

$$P = 1 - \alpha$$

$$t_{a,v} = t(a, n_1 + n_2 - 2)$$

$$s_e^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\operatorname{var} X_1 + \operatorname{var} X_2}{n_1 + n_2 - 2},$$

$$s_r^2 = s_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

H_0: $m_1 - m_2 = a$ $\quad$ (H_1: $m_1 - m_2 \neq a$)

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2 - a}{s_r}$$

H_0: $m_1 - m_2 = 0$ $\quad$ (H_1: $m_1 - m_2 \neq 0$)

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{s_r}$$

Jeżeli zajdzie nierówność $\left| t_{emp} \right| \geq t(a, n_1 + n_2 - 2)$, to hipotezę H_0 należy odrzucić na korzyść hipotezy alternatywnej H_1. Natomiast, gdy prawdziwa będzie nierówność przeciwna, tzn. $\left| t_{emp} \right| < t(a, n_1 + n_2 - 2)$, to nie mamy podstaw do odrzucenia hipotezy H_0.

Przykład:

Badano średnie oceny ze wszystkich wykładanych przedmiotów, po pierwszym semestrze studiów, studentów dwóch wydziałów SGGW. Dla losowo wybranych osób uzyskano wyniki:

x_{1i}	3,12	2,73	3,17	3,44	3,87	4,55	3,94	-
x_{2i}	2,47	2,42	3,07	2,57	3,11	3,39	3,81	3,55

Czy można stwierdzić, na podstawie tych danych, że średnie dla obu porównywanych wydziałów są jednakowe (nie różnią się istotnie) ?

Zakładamy, że rozkłady dwu badanych, niezależnych populacji są normalne o równych wariancjach.

Dla dwu prób otrzymujemy następujące parametry:

$$n_1 = 7, \qquad \bar{x}_1 = 3.55, \qquad s_1^2 = 0.378$$

$$n_2 = 8, \qquad \bar{x}_2 = 3.05, \qquad s_2^2 = 0.273$$

Sposób I:

Posłużymy się przedziałem ufności dla różnicy między nieznanymi średnimi m_1 i m_2.

$$\{(\bar{x}_1 - \bar{x}_2) - t(a, n_1 + n_2 - 2)s_r ; (\bar{x}_1 - \bar{x}_2) + t(a, n_1 + n_2 - 2)s_r\}$$

$$P = 1 - a$$

Dla naszych danych t(0.05;13) = 2.16, zaś błąd różnicy średnich s_r = 0,293.

Stąd

$$m_1 - m_2 \in (-0.137; 1.131) \qquad P = 0.95$$

Ponieważ wyznaczony przedział ufności zawiera zero, możemy uznać, że różnica między średnimi ocenami w porównywanych wydziałach **nie różni się istotnie** (średnie oceny są równe). Jest to wniosek o pewności 0.95, inaczej przedział o ufności 95%.

Sposób II:

Formułujemy odpowiednią hipotezę statystyczną i testujemy ją:

$$H_0: m_1 - m_2 = 0 \qquad (H_1: m_1 - m_2 \neq 0)$$

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{s_r} = \frac{3.55 - 3.05}{0.293} = 1.706$$

Ponieważ $|t_{emp}| < 2.16$, to na poziomie istotności 5% nie ma podstaw do odrzucenia hipotezy o tym, że średnie populacyjne są takie same. W dalszych postępowaniach możemy zakładać, że średnie oceny analizowanych wydziałów nie różnią się (obserwowane średnie z prób **nie różnią się istotnie)**.

Analiza porównawcza dwu średnich populacyjnych w przypadku **zależnych** populacji przebiega zupełnie inaczej (tzw. testy nieparametryczne).

Ocena ilorazu dwóch odchyleń standardowych (wariancji): σ_1/σ_2

Weryfikacja hipotezy zerowej o równości dwu średnich populacyjnych wymaga spełnienia założenia, że badane cechy (populacje) mają rozkłady normalne o takiej samej wariancji.

Sprawdzenie tego założenia oparte jest o rozkład Fishera-Snedecora.

Założenia:

1. $X_1 \sim N(m_1, s_1^2)$, $X_2 \sim N(m_2, s_2^2)$,
2. wartości parametrów *nie są znane*
3. X_1, X_2 *są niezależne.*

Formułujemy przypuszczenie o równości wariancji:

$$H_0 : s_1^2 = s_2^2 \quad , \quad H_1 : s_1^2 > s_2^2$$

(wybór jednostronnej hipotezy alternatywnej jest konsekwencją konstrukcji tablic statystycznych)

Funkcja testowa ma postać:

$$F_{emp} = \frac{s_1^2}{s_2^2}$$

przy czym wartość ilorazu wariancji próbkowych powinna być ułamkiem niewłaściwym, to znaczy nie mniejsza od jedności (większa wariancja do licznika — oznacza to, że próba 1 to próba o większej wariancji) .

Z tablic rozkładu F Fishera-Snedecora odczytujemy wartość $F_{\alpha, u, v}$ i jeśli $F_{emp} > F_{\alpha, u, v}$, to hipotezę zerową odrzucamy. W przeciwnym razie nie ma podstaw do odrzucenia testowanego przypuszczenia.

Stopnie swobody liczymy według znanych formuł: $u = n_1 - 1$, a $v = n_2 - 1$.

Przykład:

Sprawdźmy założenie o równości wariancji dla wcześniejszego przykładu dotyczącego porównania średnich ocen na dwóch wydziałach SGGW. Otrzymaliśmy następujące parametry dla prób:

$$n_1 = 7, \qquad s_1^2 = 0.378$$

$$n_2 = 8, \qquad s_2^2 = 0.273$$

Formułujemy hipotezę zerową:

$$H_0 : s_1^2 = s_2^2 \quad , \quad H_1 : s_1^2 > s_2^2$$

i testujemy ją

$$F_{emp} = \frac{s_1^2}{s_2^2} = \frac{0.378}{0.273} = 1.385$$

Ponieważ $F_{\alpha,\ u,\ v} = F_{0.05,\ 6,\ 7} = 3.866$, to na poziomie istotności 5% nie mamy podstaw do odrzucenia hipotezy zerowej o równości wariancji w tych populacjach. Oznacza to, że założenie o równości wariancji badanych populacji jest spełnione.

Porównanie dwóch rozkładów dwupunktowych

Założenia:

1. $X_1 \sim D(p_1), \quad X_2 \sim D(p_2),$

2. *X_1, X_2 są niezależne.*

Ocena różnicy między parametrami ($p_1 - p_2$)

$p_1 - p_2$

$\hat{p}_1 - \hat{p}_2$

$\left(\hat{p}_1 - \hat{p}_2 \pm u_{1-a/2}\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}\right)$

H_0: $p_1 - p_2 = a$

H_1: $p_1 - p_2 \neq a$

gdzie: $\hat{p}_1 = \frac{k_1}{n_1}, \quad \hat{p}_2 = \frac{k_2}{n_2}, \quad \hat{p} = \frac{k_1 + k_2}{n_1 + n_2}$

zazwyczaj interesuje nas równość frakcji, H_0: $p_1 - p_2 = 0$ (tzn. H_0: $p_1 = p_2$) i H_1: $p_1 - p_2 \neq 0$ (tzn. H_0: $p_1 \neq p_2$)

$\mathbf{H_0: p_1=p_2}$ $(H_1: \mathbf{p_1 \neq p_2})$

$$z_{emp} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Jeśli $|z_{emp}| \geq u_{1-\alpha/2}$ $(=t_{\alpha,\infty})$ to na poziomie istotności α hipotezę H_0 należy odrzucić na korzyść H_1. W przeciwnym przypadku ($|z_{emp}| < u_{1-\alpha/2}$) nie mamy podstaw do odrzucenia hipotezy H_0.

Przykład:

Zweryfikujmy na poziomie $\alpha = 0{,}05$ przypuszczenie, że wskaźniki wyposażenia studentów dwu uczelni w komputery mobilne są porównywalne, jeśli dla uczelni I w losowej próbie 800 studentów fakt posiadania komputera zadeklarowało 350 osób. Dla uczelni II odpowiednie wyniki były następujące: 1000 i 400.

$$\hat{p}_1 = \frac{k_1}{n_1} = \frac{350}{800} = 0.4375$$

$$\hat{p}_2 = \frac{k_2}{n_2} = \frac{400}{1000} = 0.4$$

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2} = \frac{350 + 400}{1800} = 0.417$$

Sposób I:

$$\left(\hat{p}_1 - \hat{p}_2 \pm u_{1-a/2}\sqrt{\hat{p}(1-\hat{p})(\tfrac{1}{n_1} + \tfrac{1}{n_2})}\right)$$

na poziomie ufności 1–α

$$\left(0{,}4375 - 0{,}4 \pm 1{,}96\sqrt{0{,}417 \cdot 0{,}583 \cdot (\tfrac{1}{800} + \tfrac{1}{1000})}\right)$$

$$\left(0{,}0375 \pm 1{,}96 \cdot 0{,}023\right)$$

$$p_1 - p_2 \in \left(-0.008;\, 0.083\right) \quad P = 0{,}95$$

Badane wskaźniki obu uczelni nie różnią się istotnie. Zaufanie do tego wniosku jest na poziomie 0.95.

Sposób II:

$$\mathbf{H_0: p_1=p_2} \qquad (H_1: \mathbf{p_1 \neq p_2})$$

$$z_{emp} = \frac{0{,}4375 - 0{,}4}{\sqrt{0{,}417(1-0{,}417)\left(\frac{1}{800}+\frac{1}{1000}\right)}} = 1{,}603$$

Ponieważ $t_{\alpha,\infty}=1{,}96$ (czyli $|z_{emp}| < t_{\alpha,\infty}$), to na poziomie istotności **5%** nie mamy podstaw do **odrzucenia** badanej hipotezy (H_0) o tym, że **frakcje** studentów wyposażonych w komputery mobilne **są takie same**.

Przykłady dla hipotez jednostronnych

Podejrzewamy, że na skutek upadku maszyna B odmierza dawkę mniej precyzyjnie niż maszyna A (precyzja –> miarą zróżnicowania jest wariancja). Zatem chcemy porównać wariancje w 2 populacjach — populacji dawek odmierzonych maszyną A i populacji dawek odmierzonych maszyną B. Jaki jest problem merytoryczny?

- Czy **zróżnicowanie** (wariancja) dawek odmierzanych maszyną B **jest większe**

Interesuje nas, czy populacja B ma większą wariancję niż populacja A. Jaką stawiamy hipotezę?

- H0: **wariancja** wielkości dawek odmierzanych maszyną A **jest taka sama** jak wariancja dawek odmierzanych maszyną B
- $H_0: \sigma_A^2 = \sigma_B^2$

(zakładamy fakt równości wariancji, i patrzymy, czy obserwacje nie przeczą temu założeniu)

Jaka jest hipoteza alternatywna?

$H_0: \sigma_A^2 < \sigma_B^2$

Opony nowszej generacji mają (wg producenta) zmniejszoną ścieralność (można na nich przejechać dłuższą trasę).

Problem merytoryczny:

- Czy przeciętna (średnia) odległość możliwa do przejechania na oponach nowszej generacji jest większa

Hipoteza zerowa (statystyczna)

- H0: **średnia** odległość możliwa do przejechania na oponach I generacji **jest taka sama** jak dla opon II generacji
- $H_0: \mu_I = \mu_{II}$

Hipoteza alternatywna

- H1: **średnia** odległość możliwa do przejechania na oponach I generacji **jest mniejsza** niż dla opon II generacji
- $H_0: \mu_I < \mu_{II}$