Statystyczna analiza danych SAD-2020/2021

Wykład 10

Parametryczne testy istotności

Schemat postępowania

- $\alpha \in (0,1)$ – poziom istotności testu, mała liczba rzędu 0,01; 0,05; 0,1, ….

- Sformułowanie założeń o rozkładzie cechy w populacji (wybór modelu):

➢ Cecha X ma rozkład prawdopodobieństwa zależny od nieznanego parametru θ

➢ $X_1, X_2, \dots, X_n$ – prosta próba losowa, $X_j \sim X$

➢ $x_1, x_2, \dots, x_n$ – próbka

Parametryczne testy istotności

1. $H_0: \theta = \theta_0$ przeciwko $H_1: \theta > \theta_0$

lub: $H_1: \theta < \theta_0$ lub: $H_1: \theta \neq \theta_0$

2. Statystyka testowa $G := G(X_1, X_2, \ldots, X_n, \theta_0)$ ma znany rozkład, jeśli hipoteza zerowa prawdziwa

3. Zbiór krytyczny C = podzbiór zbioru wartości statystyki testowej taki, że

$$P(G \in C | H_0 - \text{prawdziwa}) = \alpha$$

4. Obliczenie wartości statystyki testowej

$$G_{obs} = G(x_1, x_2, \ldots, x_n, \theta_0)$$

Parametryczne testy istotności

5. Podjęcie decyzji na podstawie 4. według reguły:

- Jeśli $G_{obs} \in C$, to przyjmujemy H_1 na poziomie istotności α (α – prawdopodobieństwo błędnej decyzji)

- Jeśli $G_{obs} \notin C$, to nie mamy podstaw do odrzucenia H_0 (przyjęcia H_1) na poziomie istotności α

Uwaga: Modele parametrycznych testów istotności są na ftp/public/elaw/Informatyka dzienne2021 w testowanie hipotez – modele.pdf

Testy o różnicy wartości średnich dwóch rozkładów normalnych (znane wariancje)

Niech $X_1, X_2, \ldots, X_{n_1}$ oraz $Y_1, Y_2, \ldots, Y_{n_2}$ będą dwiema niezależnymi prostymi próbami losowymi z rozkładów normalnych $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$, odpowiednio.

$H_0 : \mu_1 = \mu_2$ lub równoważnie $H_0 : \mu_1 - \mu_2 = 0$.

Statystyka testowa

Statystka $\bar{X} - \bar{Y}$ ma rozkład **normalny** o wartości średniej $\mu_1 - \mu_2$ i wariancji

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Stąd $$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \sim N(0,1).$$

Testy o różnicy wartości średnich dwóch rozkładów normalnych (znane wariancje)

a) $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 > 0$.

Jeśli ***H₀*** **prawdziwa**, to $Z = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \sim N(0,1).$

Przyjmujemy $C = \{z : z \geq z_{1-\alpha}\}$ **= zbiór krytyczny** testu hipotezy ***H₀***

przeciw ***H₁*** na **poziomie istotności** α, gdzie

$$P_{H_0}(Z \in C) = P_{H_0}(Z \geq z_{1-\alpha}) = \alpha,$$

$z_{1-\alpha}$ = **kwantyl rzędu 1-** α **rozkładu** $N(0,1)$.

Testy o różnicy wartości średnich dwóch rozkładów normalnych (znane wariancje)

b) $\boxed{H_0 : \mu_1 - \mu_2 = 0}$, $\boxed{H_1 : \mu_1 - \mu_2 < 0}$.

Przyjmujemy ***C*** **=** $\{z : z \leq z_\alpha\}$ **= zbiór krytyczny.**

(c) $\boxed{H_0 : \mu_1 - \mu_2 = 0}$, $\boxed{H_1 : \mu_1 - \mu_2 \neq 0}$

Przyjmujemy ***C*** = $\{z : |z| \geq z_{1-\alpha/2}\}$ = **zbiór krytyczny**

Testy o różnicy wartości średnich dwóch rozkładów normalnych (znane wariancje)

Przykład. Średnia waga losowo wybranych 15 Europejczyków wyniosła $\bar{x}$ = 154 (funty), podczas gdy dla próbki 18 Amerykanów otrzymano $\bar{y}$ = 162 (funty).
Z poprzednich badań wiadomo, że wariancje wag losowo wybranego Europejczyka i Amerykanina wynoszą, odpowiednio: $\sigma_1^2 = 100$ i $\sigma_2^2 = 169$.
Czy można twierdzić, że średnie wagi w populacji Europejczyków i Amerykanów są różne? Przyjąć α = 0,05 oraz rozkład normalny wag.
Rozwiązanie:

1. $H_0 : \mu_1 - \mu_2 = 0$. $H_1 : \mu_1 - \mu_2 \neq 0$

2. **Statystyka testowa:** $$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Testy o różnicy wartości średnich dwóch rozkładów normalnych (znane wariancje)

3. $\alpha = 0{,}05,\ 1-\alpha/2 = 0{,}975,\ z_{0{,}975} = 1{,}96.$

Zbiór krytyczny C = $\{z : |z| \geq 1{,}96\}.$

4. Mamy $\bar{x}$ = 154, $\bar{y}$ =162, $\sigma_1^2 = 100,\ \sigma_2^2 = 169,\ n_1 = 15,\ n_2 = 18.$

Stąd $z = \dfrac{154-162}{\sqrt{100/15+169/18}} = \dfrac{-8}{\sqrt{16{,}056}} = -2.$

5. $|-2| = 2 \geq 1{,}96$, więc odrzucamy H_0.

Odpowiedź: Na poziomie istotności $\alpha = 0{,}05$ stwierdzamy, że średnia waga Europejczyka różni się od średniej wagi Amerykanina, przy czym dane sugerują, że średnio Amerykanie ważą więcej niż Europejczycy.

Test o różnicy wartości średnich dwóch rozkładów normalnych (nieznane równe wariancje)

Założenie dodatkowe: $\sigma_1 = \sigma_2 = \sigma$, σ **- nieznane.**

$\boxed{H_0 : \mu_1 = \mu_2}$, lub równoważnie $\boxed{H_0 : \mu_1 - \mu_2 = 0}$.

Statystyka testowa: $\boxed{Z = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}} \boxed{= \dfrac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/n_1 + 1/n_2}}}$

Jeśli $\boxed{H_0}$ **prawdziwa**, to $Z \sim N(0,1)$.

Wiadomo, że $\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$, oraz

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$$ -

nieobciążone estymatory σ^2.

Estymatorem nieobciążonym σ^2, opartym na dwu próbach łącznie, jest statystyka

$$\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

Wówczas we wzorze na *Z* podstawiając $S_p = \sqrt{S_p^2}$ zamiast σ otrzymujemy statystykę

$$\boxed{T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \sim \boxed{t_{n_1 + n_2 - 2}}.$$

Test o różnicy wartości średnich dwóch rozkładów normalnych (nieznane równe wariancje)

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Dla trzech przypadków możliwych hipotez alternatywnych (a), (b), (c) z modelu poprzedniego mamy analogiczne obszary krytyczne, przy czym kwantyle rozkładu $N(0,1)$ zastępujemy kwantylami rozkładu $t_{n_1+n_2-2}$.

Test o różnicy wartości średnich dwóch rozkładów normalnych (nieznane równe wariancje)

Przykład. Klasyczne tranzystory domieszkowane złotem (występujące w układach scalonych) mają tzw. czas magazynowania ładunku rzędu 7 ns. Producent ma nadzieję, że pewna zmiana technologii zmniejszyła czas magazynowania. Producent chciałby przetestować hipotezę $H_0 : \mu_1 = \mu_2$ przeciw $H_1 : \mu_1 > \mu_2$, gdzie μ_1 oznacza średni czas magazynowania przy starej technologii a μ_2 przy nowej technologii. Z poprzednich badań wiadomo, że obie technologie dają w przybliżeniu normalne rozkłady czasu magazynowania, oraz że odchylenia standardowe obu rozkładów są takie same. Producent pobrał 2 niezależne 50 elementowe próbki tranzystorów, produkowanych starą i nowa technologią.
Średnie czasy magazynowania dla obu próbek wyniosły

$$\bar{x} = 6{,}6, \quad \bar{y} = 6{,}3 \text{ oraz } s_p = 0{,}5.$$

Test o różnicy wartości średnich dwóch rozkładów normalnych (nieznane równe wariancje)

Statystyka testowa:

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{98}, \text{ jeśli } H_0 \text{ prawdziwa}$$

Wartość statystyki testowej:

$$t = \frac{6{,}6 - 6{,}3}{0{,}5 \sqrt{\frac{1}{50} + \frac{1}{50}}} = 3{,}0$$

Test o różnicy wartości średnich dwóch rozkładów normalnych (nieznane równe wariancje)

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 > \mu_2.$$

Stąd **obszar krytyczny**

$$C = \{t : t \geq t_{1-\alpha,98} \cong z_{1-\alpha}\} =< z_{1-\alpha}, \infty)$$

oraz **p-wartość** testu wynosi $P_{H_0}(T \geq 3{,}0) = 0{,}002.$ Zatem, można przyjąć, że nowa technologia zmniejszyła średni czas magazynowania ładunku.

Test o różnicy wartości średnich rozkładów brzegowych (dane „sparowane”)

Niech $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ będzie prostą próbą losową z rozkładu dwuwymiarowego. Niech $D_i = X_i - Y_i, i = 1,..,n,$ tworzą prostą próbę losową z rozkładu normalnego o nieznanej średniej μ_D.

Hipoteza zerowa: $\boxed{H_0 : \mu_D = 0}$,

Możliwe hipotezy alternatywne:

$\boxed{H_1 : \mu_D > 0}$ $\boxed{H_1 : \mu_D < 0}$ $\boxed{H_1 : \mu_D \neq 0}$.

Statystyka testowa: $\boxed{T = \dfrac{\overline{D}}{S_D / \sqrt{n}}}$.

Test o różnicy wartości średnich rozkładów brzegowych

Jeśli H_0 **prawdziwa**, to $T \sim t_{n-1}$

Zatem, obszary krytyczne takie same jak przy testowaniu hipotez o wartości średniej jednej populacji normalnej przy nieznanym odchyleniu standardowym.

Test o różnicy wartości średnich dla danych „sparowanych”

Przykład. Zmierzono ciśnienie tętnicze wśród losowo wybranej grupy chorych na pewną chorobę przed i po podaniu takiego samego leku każdemu z pacjentów. Otrzymano następujące wyniki:

Pacjent:	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Przed :	210	180	260	270	190	250	180
Po :	180	160	220	260	200	230	180

Czy można twierdzić, na poziomie istotności 0,05, że lek powoduje zmniejszenie wartości średniej ciśnienia?(podać odpowiednie założenia).

1 . $\boxed{H_0 : \mu_1 = \mu_2} \equiv \boxed{H_0 : \mu_D = \mu_1 - \mu_2 = 0}$

$\boxed{H_1 : \mu_1 > \mu_2} \equiv \boxed{H_1 : \mu_D = \mu_1 - \mu_2 > 0}$

3. Statystyka testowa: $\boxed{T = \dfrac{\overline{D}}{S_D / \sqrt{n}}}$.

Test o różnicy wartości średnich dla danych „sparowanych”

4. d_i: 30, 20, 40, 10, -10, 20, 0, $\bar{d} = 15{,}7$, s_D = 15,9, n = 7.

$$t = \frac{15{,}7}{15{,}9/\sqrt{7}} = 2{,}24$$

5. $\alpha = 0{,}05$, $1-\alpha = 0.95$, $n-1 = 7-1 = 6$,

$$t_{0{,}95,6} = 1{,}94$$

6. 2,24 >1,94, więc odrzucamy hipotezę zerową.

Odpowiedź. Można twierdzić, że lek obniżył wartość średnią ciśnienia w populacji pacjentów, na poziomie istotności 0,05.

Test dla proporcji (wskaźnika struktury)

Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu Bernoulli'ego o nieznanym parametrze *p*. Wówczas $\mu = E(X_1) = p$, $\sigma^2 = p(1-p)$.

Np. gdy *p* jest proporcją obiektów populacji mających pewną własność, przyjmujemy X_i = 1 (0) gdy wylosowany obiekt posiada (nie posiada) tę własność. Niech $\hat{p} = \overline{X}$ = częstość = proporcja elementów próby o danej własności. Z CTG dla dostatecznie dużego *n* zmienna losowa

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

ma rozkład bliski rozkładowi standardowemu normalnemu $N(0,1)$. (musi zachodzić $np \geq 5, n(1-p) \geq 5$).

Test dla proporcji

Hipoteza zerowa: $H_0 : p = p_0$

Statystyka testowa

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \text{bliski } N(0,1),$$

jeśli hipoteza H_0 jest prawdziwa. Możliwe sytuacje:

- $H_1 : p > p_0, \quad C = \{z : z \geq z_{1-\alpha}\}$
- $H_1 : p < p_0, \quad C = \{z : z \leq -z_{1-\alpha}\}$
- $H_1 : p \neq p_0, \quad C = \{z : |z| \geq z_{1-\alpha/2}\}$

Test dla proporcji

Przykład. Przypuszczamy, że proporcja samochodów w Warszawie używających gazu jako paliwa jest mniejsza niż 0,15. W próbie 200 losowo samochodów 21 było samochodami na gaz. Czy te dane potwierdzają nasze przypuszczenie, przy poziomie istotności 0,05 ?

1. $H_0: \ p = 0{,}15 \ , \quad H_1 : p < 0{,}15$

2. Statystyka testowa

$$Z = \frac{\hat{p} - 0{,}15}{\sqrt{\frac{0{,}15 \times 0{,}85}{200}}} \sim \text{bliski } N(0,1),$$ jeśli hipoteza H_0 jest prawdziwa.

Test dla proporcji

3. Wartość statystyki testowej dla próbki:

$$Z = \frac{21/200 - 0{,}15}{\sqrt{\frac{0{,}15 \times 0{,}85}{200}}} = -1.79$$

4. Kwantyl $z_{0,95} = 1{,}64$

5. Zbiór krytyczny $C = \{z : z \leq -1{,}64\}$

6. $-1{,}79 \in C$, więc stwierdzamy, że proporcja samochodów na gaz jest mniejsza niż 0,15, przyjmując poziom istotności 0,05 (0,05 = prawdopodobieństwo, że nasza decyzja jest błędna)

Test dla różnicy proporcji dwóch populacji

Niech $X_1, X_2, \ldots, X_{n_1}$ oraz $Y_1, Y_2, \ldots, Y_{n_2}$ będą dwiema niezależnymi prostymi próbami losowymi z dwu populacji mających rozkłady Bernoulli'ego o nieznanych parametrach p_1, p_2 , odpowiednio.

Niech K_1 oraz K_2 będą liczbami elementów próby X' ów oraz Y' ów o wartościach 1, odpowiednio.
Estymatory proporcji p_1, p_2:

$$\hat{p}_1 = \frac{K_1}{n_1}, \qquad \hat{p}_2 = \frac{K_2}{n_2}, \text{ odpowiednio}$$

Estymator *p,* jeśli $H_0: p_1 = p_2 = p$ jest prawdziwa:

$$\hat{p} = \frac{K_1 + K_2}{n_1 + n_2}$$

Test dla różnicy proporcji dwóch populacji

Hipoteza zerowa

$$H_0: p_1 = p_2$$

Statystyka testowa

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

ma rozkład bliski N(0,1), jeśli H_0 prawdziwa

Test dla różnicy proporcji dwóch populacji

oraz

$$n_1\hat{p}_1 \geq 5,\ \ n_1(1-\hat{p}_1) \geq 5,\ n_2\hat{p}_2 \geq 5,\ \ n_2(1-\hat{p}_2) \geq 5$$

Możliwe sytuacje

- $H_1: p_1 > p_2$, wówczas $C = \{z: z \geq z_{1-\alpha}\}$
- $H_1: p_1 < p_2$, wówczas $C = \{z: z \leq -z_{1-\alpha}\}$
- $H_1: p_1 \neq p_2$, wówczas $C = \{z: |z| \geq z_{1-\alpha/2}\}$

Test dla różnicy proporcji dwóch populacji

Przykład. Porównywano monitory firmy A i B. Spośród 200-tu monitorów firmy A 10 wymagało naprawy w okresie gwarancji, natomiast spośród 150-ciu monitorów firmy B 12 wymagało naprawy w okresie gwarancji. Czy można twierdzić, że prawdopodobieństwo awarii monitora firmy A jest mniejsze niż prawdopodobieństwo awarii monitora firmy B. Przyjąć poziom istotności 0,06.

1. $H_0: p_1 = p_2$, $H_1: p_1 < p_2$

2. Statystyka testowa

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

Test dla różnicy proporcji dwóch populacji

3. $$z = \frac{\frac{10}{200} - \frac{12}{150}}{\sqrt{\left(\frac{1}{200}+\frac{1}{150}\right)\frac{10+12}{200+150}\left(1-\frac{22}{350}\right)}} =$$

$$\frac{0{,}05-0{,}08}{\sqrt{\frac{200+150}{200\cdot 150}\cdot 0{,}063\cdot 0{,}932}} = -1{,}14$$

4. $1-\alpha = 0{,}94$, $z_{0,94} = 1{,}55$, $C = (-\infty, -1{,}55]$
5. $-1{,}14$ nie należy do C
6. Na poziomie istotności 0,06 nie można odrzucić hipotezy, że monitory obu firm mają jednakowe prawdopodobieństwo awarii w okresie gwarancji

Test zgodności chi-kwadrat

Założenia

1. Badana cecha jednostek populacji może przyjmować *k* różnych wartości (może należeć do *k* różnych klas, kategorii): c_1, c_2,…, c_k. Niech zmienna losowa *X* oznacza kategorię (klasę) losowo wybranej jednostki.
2. $H_0: P(X = c_1) = p_1,\ P\ (X = c_2) = p_2, \dots,\ P(X = c_k) = p_k$.
3. Dla próby losowej cech *n* losowo wybranych jednostek populacji niech $N_1, N_2, \dots, N_k$ oznaczają liczności jednostek o cechach c_1, c_2,…, c_k, odpowiednio.
4. Jeśli hipoteza zerowa jest prawdziwa to oczekiwane liczności wynoszą:

$$EN_1 = np_1,\ EN_2 = np_2, \dots,\ EN_k = np_k.$$

Test zgodności chi-kwadrat

5. Odstępstwo empirycznych liczności (z próby) od oczekiwanych liczności jest mierzone za pomocą statystyki chi-kwadrat χ^2 postaci:

$$\frac{(N_1-EN_1)^2}{EN_1}+\frac{(N_2-EN_2)^2}{EN_2}+\cdots+\frac{(N_k-EN_k)^2}{EN_k}.$$

6. Jeśli wszystkie oczekiwane liczności są nie mniejsze niż 5, tzn $EN_j \geq 5$, $j = 1,2,..,k$, to rozkład χ^2 można przybliżyć rozkładem chi-kwadrat. Jest k kategorii więc liczba stopni swobody rozkładu chi-kwadrat wynosi wynosi $k - 1$.

Test zgodności chi-kwadrat

7. Jeśli H_0 jest prawdziwa, to odstępstwo empirycznych liczności od oczekiwanych liczności powinno być małe. Stąd wartości statystyki χ^2 też powinny być małe. Z kolei jeśli występuje duża rozbieżność pomiędzy obserwowanymi licznościami kategorii a "teoretycznymi", to wątpimy o prawdziwości H_0 w przypadku dużych wartości statystyki χ^2. Stąd zbiór krytyczny ma postać:

$$C = \left\{\chi^2 : \chi^2 \geq \chi^2_{1-\alpha,k-1}\right\} = [\chi^2_{1-\alpha,k-1}, \infty).$$

8. Reguła decyzyjna: Odrzucenie H_0 , jeśli obliczona wartość $\chi^2 \geq \chi^2_{1-\alpha,k-1}$

Test zgodności chi-kwadrat

Przykład. Przypuszcza się, że proporcje ludzi z grupami krwi: A, B, AB, i 0 wynoszą, odpowiednio: 0.4, 0.2, 0.1, 0.3. Wśród 400-tu losowo wybranych osób liczby osób o powyższych grupach krwi wyniosły: 148, 96, 50, 106. Czy na poziomie istotności 5% można zaprzeczyć powyższemu przypuszczeniu?

Rozwiązanie.

$$H_0: p_A = 0.4,\ p_B = 0.2,\ p_{AB} = 0.1,\ p_0 = 0.3.$$

Test zgodności chi-kwadrat

Obliczenie wartości χ^2:

Grupa krwi	Liczności z próbki N	Średnie liczności EN	$(N - EN)$	$(N - EN)^2$	$(N - EN)^2/EN$
A	148	160	- 12	144	0.90
B	96	80	16	256	3.20
AB	50	40	10	196	2.50
0	106	120	-14	100	1.63
suma	400	400	0		**8.23**

Test zgodności chi-kwadrat

Liczba stopni swobody: k – 1 = 4 – 1 = 3

Poziom istotności testu α = 0,05, stąd 1 – α = 0,95

Kwantyl $\chi^2_{0.95,3}$ = 7,81.

Wartość statystyki chi-kwadrat 8,23 > 7,81, więc odrzucamy hipotezę zerową.

Test niezależności cech

- **Cel**: testowanie hipotezy, że dwie cechy jednostek populacji są niezależne.

- **Przykłady**:

Grupa krwi i kolor oczu
Wiek i zapatrywania polityczne
Kolor oczu i kolor włosów
Picie alkoholu i palenie papiersów
Dochód i wykształcenie
Podatki i PKB
Niezawodność systemu i producent

Test niezależności cech

Tablica kontyngencyjna

	d_1	d_2		d_j		d_r	
c_1	n_{11}	n_{12}		n_{1i}		n_{1r}	$n_{1\bullet}$
c_2	n_{21}	n_{22}		n_{2i}		n_{2r}	$n_{2\bullet}$
c_i	n_{i1}	n_{i2}		n_{ij}		n_{ir}	$n_{i\bullet}$
c_k	n_{k1}	n_{k2}		n_{kj}		n_{kr}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet j}$		$n_{\bullet r}$	n

Test niezależności cech

Założenia oraz test

1. Jednostka populacji scharakteryzowana jest parą cech (atrybutów). Niech (X,Y) będzie parą atrybutów wybranej losowo jednostki populacji. Możliwe wartości X należą do k różnych klas (kategorii): $c_1, c_2,\ldots, c_k$ możliwe wartości cechy Y należą do r różnych klas (kategorii): $d_1, d_2,\ldots, d_r$.

2. H_0 : X, Y są niezależnymi zmiennymi losowymi:

$$H_0: P(X = c_1, Y = d_1) = P(X = c_1)P(Y = d_1),\ldots,$$

$$P(X = c_k, Y = d_r) = P(X = c_k)P(Y = d_r).$$

W skrócie $H_0: p_{ij} = p_{i.}p_{.j}, \quad i = 1,2 \ldots, k; j = 1,2, \ldots, r$

Test niezależności cech

Założenia oraz test (kont.)

3. Niech N_{ij} będzie liczbą elementów prostej próby losowej o liczności n(próbki) z tej populacji, dla których cechą pierwszą jest klasa *i,* a drugą klasa *j.* Niech dla próbki:
   n_{ij} = liczba elementów próbki o charakterystykach
   $$c_i, d_j, i = 1,2, \dots, k, j = 1,2 \dots, r.$$
4. Jeśli H_0 prawdziwa, to oczekiwane liczby obserwacji o charakterystykach (c_i, d_j) wynoszą:

$$np_{ij} = np_{i\cdot}p_{\cdot j}\text{, gdzie } p_{i\cdot} = P(X = c_i),\ p_{\cdot j} = P\left(Y = d_j\right).$$

Uzasadnienie: $N_{ij} \sim Bin\left(n, p_{ij}\right) \Longrightarrow E\left(N_{ij}\right) = np_{i\cdot}p_{\cdot j}$,

$$X, Y - niezależne \Longrightarrow \quad p_{ij} = p_{i\cdot}p_{\cdot j}$$

Test niezależności cech

5. Odstępstwo empirycznych liczebności klas N_{ij} od oczekiwanych liczebności klas $E(N_{ij})$, przy założeniu, że hipoteza zerowa jest prawdziwa, wyraża statystyka chi-kwadrat:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{\left(N_{ij} - \widehat{N_{ij}}\right)^2}{\widehat{N_{ij}}},$$

gdzie $\widehat{N_{ij}} = \frac{N_{i.}N_{.j}}{n}$ jest estymatorem $E(N_{ij})$.

Uzasadnienie: $E(N_{ij}) = np_{i.}p_{.j} \Longrightarrow$

$$E\left(\widehat{N_{ij}}\right) = n\widehat{p_{i.}} \cdot \widehat{p_{.j}} = n\frac{N_{i.}}{n}\frac{N_{.j}}{n} = \widehat{N_{ij}}$$

$N_{i.}, N_{.j}$ – liczby elementów próby, dla których, odpowiednio, cecha X ma i-tą wartość, a cecha Y j-tą

Test niezależności cech

6. Jeśli wszystkie $\widehat{N}_{ij} \geq 5$, to można przyjąć, że rozkład statystyki chi-kwadrat jest bliski rozkładowi chi-kwadrat o liczbie stopni swobody $(k-1)(r-1)$.

7. Jeśli H_0 jest prawdziwa, to odstępstwo empirycznych liczności od estymatorów oczekiwanych liczności powinno być małe. Stąd wartości statystyki chi-kwadrat też powinny być małe. Z kolei, jeśli występuje duża rozbieżność pomiędzy obserwowanymi licznościami kategorii a estymatorami "teoretycznych" liczności, to wątpimy o prawdziwości H_0 w przypadku dużych wartości statystyki chi-kwadrat. Stąd zbiór krytyczny ma postać:

$$C = [\chi^2_{1-\alpha,(k-1)(r-1)}, \infty)$$

Test niezależności cech

8. **Reguła decyzyjna**: Odrzucenie hipotezy o niezależności cech, tzn. stwierdzenie zależności cech, na poziomie istotności α, jeśli

$$\chi^2_{obs} \geq \chi^2_{1-\alpha,(k-1)(r-1)}.$$

Jeśli zachodzi nierówność przeciwna, to nie możemy odrzucić hipotezy, że cechy X, Y są niezależne.

Test niezależności cech

Przykład Na pewnej uczelni technicznej mającej 3 wydziały A,B,C przeprowadzono egzamin semestralny ze statystyki. Niech X oznacza przynależność losowo wybranego studenta do wydziału (1 = A, 2 = B, 3 = C), a wartość Y wynosi 1, jeśli student zdał egzamin, 0 w przypadku przeciwnym.

Test niezależności cech

Wyniki badania

y / *x*	1	0	
1	350	50	400
2	450	150	600
3	200	100	300
	1000	300	1300

Obliczona wartość statystyki chi-kwadrat wynosi 44,2.

Niech poziom istotności $\alpha = 0{,}01$.

Liczba stopni swobody (3-1)(2-1) = 2

Kwantyl $\chi^2_{0,99} = 9{,}210$

Zbiór krytyczny $C = [9.210, \infty)$, $44{,}2 \in C$

Decyzja: Wynik egzaminu zależy od wydziału, przy założonym poziomie istotności 0,01.