

Statystyka Matematyczna

Anna Janicka

wykład XII, 23.05.2016

PORÓWNANIE WIĘCEJ NIŻ DWÓCH POPULACJI

TESTY NIEPARAMETRYCZNE

Plan na dzisiaj

1. Porównywanie więcej niż dwóch populacji
 - test jednoczynnikowej analizy wariancji (ANOVA)
 2. Testy zgodności
 - test Kołmogorowa
 - test Kołmogorowa-Smirnowa (dla dwóch próbek)
 - test Kołmogorowa-Lillieforsa
 - test zgodności chi-kwadrat
różne wersje
 3. Testy niezależności
 - test chi-kwadrat
-



Porównywanie dwóch i więcej populacji

Zastanawiamy się, czy populacje są pod pewnymi względami „takie same”:

- ☐ testy parametryczne: badamy równość konkretnych parametrów rozkładów
- ☐ testy nieparametryczne: sprawdzamy, czy rozkłady są takie same



Uwaga na notację:

$x_{\text{coś}}$ **zawsze** oznacza kwantyl rzędu coś



Testy dla więcej niż dwóch populacji

Naiwne podejście:

testowanie „parami”

Uwaga:

w tym przypadku p-stwo popełnienia błędu I rodzaju jest większe, niż założony poziom istotności!



WARSAW UNIVERSITY
Faculty of Economic Sciences

Przypadek wielu populacji

Przypuśćmy, że mamy k prób losowych

$$X_{1,1}, X_{1,2}, \dots, X_{1,n_1},$$

$$X_{2,1}, X_{2,2}, \dots, X_{2,n_2},$$

...

$$X_{k,1}, X_{k,2}, \dots, X_{k,n_k}, \text{ oraz}$$

- wszystkie $X_{i,j}$ są niezależne ($i=1, \dots, k, j=1, \dots, n_i$)
 - $X_{i,j} \sim N(m_i, \sigma^2)$
 - nie znamy $m_1, m_2, \dots, m_k$, ani σ^2
-



Test analizy wariancji (ANOVA) na poziomie istotności α

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1: \neg H_0 \quad (\text{tzn. nie wszystkie } \mu_i \text{ s\k{a} r\k{o}wne})$$

Test ilorazu wiarygodnořci ze statystyk\k{a}:

$$F = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 / (k-1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2 / (n-k)} \sim F(k-1, n-k)$$

ma obszar krytyczny

$$K^* = \{x : F(x) > F_{1-\alpha}(k-1, n-k)\}$$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}, \bar{X} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^k n_i \bar{X}_i$$

dla $k=2$ test ANOVA jest r\k{o}wnoważny testowi t dla dw\k{o}ch populacji



Test analizy wariancji – interpretacja

mamy

$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{i,j} - \bar{X})^2}_{\text{Sum of Squares (SS)}} = \underbrace{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2}_{\text{Sum of Squares Between (SSB)}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2}_{\text{Sum of Squares Within (SSW)}}$$

$\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$ – estymator wariancji międzygrupowej

$\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2$ – estymator wariancji wewnątrz grup



Test analizy wariancji – tabela

źródło zmienności	sumy kwadratów	liczba stopni swobody	wartość statystyki F
między grupami	SSB	$k-1$	–
wewnątrz grup	SSW	$n-k$	–
w sumie	SS	$n-1$	F



Test analizy wariancji – przykład

Roczne spożycie czekolady w trzech miastach: A, B, C na podstawie losowych prób $n_A = 8$, $n_B = 10$, $n_C = 9$ konsumentów. Czy średnie spożycie zależy od miasta?

	A	B	C
średnia z próby	11	10	7
wariancja z próby	3,5	2,8	3

$$\bar{X} = \frac{1}{27} (11 \cdot 8 + 10 \cdot 10 + 7 \cdot 9) = 9,3$$

$$SSB = (11 - 9,3)^2 \cdot 8 + (10 - 9,3)^2 \cdot 10 + (7 - 9,3)^2 \cdot 9 = 75,63$$

$$SSW = 3,5 \cdot 7 + 2,8 \cdot 9 + 3 \cdot 8 = 73,7$$

$$F = \frac{75,63 / 2}{73,7 / 24} \approx 12,31 \quad \text{a} \quad F_{0,99}(2,24) \approx 5,61$$



Test analizy wariancji – tabela – przykład

źródło zmienności	sumy kwadratów	liczba stopni swobody	wartość statystyki F
między grupami	75,63	2	–
wewnątrz grup	73,7	24	–
w sumie	149,33	26	12,31



Testy nieparametryczne

- ☐ Badamy, czy zmienna pochodzi z konkretnego rozkładu (testy *zgodności*).
- ☐ Badamy, czy rozkłady zmiennych są takie same
- ☐ Badamy, czy zmienne/cechy są niezależne (test niezależności)



Test zgodności Kołmogorowa

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu o dystrybuancie F .

$$H_0: F = F_0 \quad (F_0 \text{ ustalona})$$

$$H_1: \neg H_0 \quad (\text{tzn. dystrybuanta jest jakaś inna})$$

Jeśli F_0 jest ciągła, to testujemy statystyką

$$D_n = \sup_{t \in R} |F_n(t) - F_0(t)| = \max\{D_n^+, D_n^-\}$$

gdzie

$$D_n^+ = \max_{i=1, \dots, n} \left| \frac{i}{n} - F_0(x_{i:n}) \right|, \quad D_n^- = \max_{i=1, \dots, n} \left| F_0(x_{i:n}) - \frac{i-1}{n} \right|$$

zaś $F_n(t)$ – n -ta dystrybuanta empiryczna



Test zgodności Kołmogorowa – cd.

Postać testu: odrzucamy H_0 gdy:

$$D_n > c(\alpha, n)$$

dla pewnej wartości krytycznej $c(\alpha, n)$.

Tw. Przy prawdziwej H_0 rozkład statystyki D_n nie zależy od rozkładu F_0 .

Problem: Ten rozkład wymaga tablicowania, w zasadzie dla każdego n z osobna

Tw. W granicy

$$P(\sqrt{n}D_n \leq d) \xrightarrow{n \rightarrow \infty} K(d) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 d^2}$$



można stosować przybliżenie dla $n \geq 100$

Test zgodności Kołmogorowa – cd. 2

Tablica rozkładu asymptotycznego $K(d)$

$1-\alpha$	0,8	0,9	0,95	0,99
kwantyl $K(d)$	1,07	1,22	1,36	1,63
$c(n, \alpha)$ dla $n \geq 100$	$1,07 / \sqrt{n}$	$1,22 / \sqrt{n}$	$1,36 / \sqrt{n}$	$1,63 / \sqrt{n}$



Test zgodności Kołmogorowa – przykład

Czy próba

0,4085 0,5267 0,3751 0,8329 0,0846

0,8306 0,6264 0,3086 0,3662 0,7952

pochodzi z rozkładu jednostajnego $U(0,1)$?



Test zgodności Kołmogorowa – przykład cd.

$X_{i:10}$	$(i-1)/10$	$i/10$	$i/10 - F(X_{i:10})$	$F(X_{i:10}) - i/10$
0,0846	0	0,1	0,0154	0,0846
0,3086	0,1	0,2	-0,1086	0,2086
0,3662	0,2	0,3	-0,0662	0,1662
0,3751	0,3	0,4	0,0249	0,0751
0,4085	0,4	0,5	0,0915	0,0085
0,5267	0,5	0,6	0,0733	0,0267
0,6264	0,6	0,7	0,0736	0,0264
0,7952	0,7	0,8	0,0048	0,0952
0,8306	0,8	0,9	0,0694	0,0306
0,8329	0,9	1	0,1671	-0,0671

$$D_n = 0,2086 \quad c(10; 0,9) = 0,369$$

→ nie ma podstaw do odrzucenia hipotezy o
jednostajności rozkładu



Test zgodności Kołmogorowa – Smirnowa

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu o dystrybuancie F , $Y_1, Y_2, \dots, Y_m$ są próbą IID z rozkładu o dystrybuancie G .

$$H_0: F = G$$

$$H_1: \neg H_0 \quad (\text{tzn. dystrybuanty są różne})$$

Jeśli F (i G) jest ciągła, to testujemy statystyką

$$D_{n,m} = \sup_{t \in R} |F_n(t) - G_m(t)|$$

gdzie $F_n(t)$ – n -ta dystrybuanta empiryczna pierwszej próbki, a $G_m(t)$ – m -ta dystrybuanta empiryczna drugiej próbki

Test zgodności Kołmogorowa – Smirnowa – cd.

Postać testu: odrzucamy H_0 gdy:

$$D_{n,m} > c(\alpha, n, m)$$

dla pewnej wartości krytycznej $c(\alpha, n, m)$.

Tw. Przy prawdziwej H_0 rozkład statystyki $D_{n,m}$ nie zależy od rozkładu F (ani G).

Tw. W granicy

$$P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leq d\right) \xrightarrow{n \rightarrow \infty, m \rightarrow \infty} K(d) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 d^2}$$

przybliżenie OK dla $n, m \geq 100$



Test zgodności Kołmogorowa – Lillieforsa

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu o dystrybuancie F .

H_0 : F jest dystrybuantą rozkładu normalnego
(o nieznanym parametrach)

H_1 : $\neg H_0$ (tzn. dystrybuanta jest jakaś inna)

Testujemy statystyką

$$D_n = \max\{D_n^+, D_n^-\}$$

gdzie $D_n^+ = \max_{i=1, \dots, n} \left| \frac{i}{n} - z_i \right|$, $D_n^- = \max_{i=1, \dots, n} \left| z_i - \frac{i-1}{n} \right|$
a

$$z_i = \Phi\left(\frac{X_{i:n} - \bar{X}}{S}\right)$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$



Test zgodności Kołmogorowa – Lillieforsa – cd.

Postać testu: odrzucamy H_0 gdy:

$$D_n > D_n(\alpha)$$

dla pewnej wartości krytycznej $D_n(\alpha)$.

Tw. Przy prawdziwej H_0 rozkład statystyki D_n nie zależy od konkretnego rozkładu norm.

Problem: Ten rozkład wymaga tablicowania i nie jest znana postać analityczna...

Stosowany zwł. dla próbek o liczebności do 30, kiedy jest lepszy niż test zgodności chi-

Test zgodności Kołmogorowa – Lillieforsa

wartości krytyczne

Sample Size N	Level of Significance for $D = \text{Max } F^*(X) - S_N(X) $				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.149	.153	.165	.180	.203
30	.131	.136	.144	.161	.187
Over 30	.736	.768	.805	.886	1.031
	$\sqrt{N}$	$\sqrt{N}$	$\sqrt{N}$	$\sqrt{N}$	$\sqrt{N}$

Źródło: H. Lilliefors

Test zgodności chi-kwadrat

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu dyskretnego o k wartościach (ozn. $1, \dots, k$).

H_0 : prawdopodobieństwa w rozkładzie X to

i	1	2	3	...	k
$P(X=i)$	p_1	p_2	p_3	...	p_k

H_1 : $\neg H_0$ (tzn. rozkład jest jakiś inny)

etykiety

Oznaczmy rezultat doświadczenia jako

i	1	2	3	...	k
N_i	N_1	N_2	N_3	...	N_k

gdzie N_i oznacza liczbę uzyskanych wyników wartości i .

$$N_i = \sum_{j=1}^n 1_{X_j=i}$$



Test zgodności chi-kwadrat – postać testu

Ogólna postać testu:

$$\chi^2 = \sum \frac{(\text{wartosc obserwowana} - \text{wartosc oczekiwana})^2}{\text{wartosc oczekiwana}}$$

u nas:

$$\chi^2 = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$$

Tw. Przy prawdziwości H_0 rozkład w/w statystyki χ^2 zmierza do rozkładu χ^2 o $k-1$ stopniach swobody $\chi^2(k-1)$ przy $n \rightarrow \infty$

Procedura: odrzucamy H_0 gdy $\chi^2 > c$, gdzie

~~$c = \chi^2_{1-\alpha}(k-1)$ jest kwantylem rzędu $1-\alpha$~~



rozkładu χ^2 o $k-1$ stopniach swobody

Test zgodności chi-kwadrat – przykład

Czy kość jest rzetelna? Na poziomie $\alpha=0,05$
 $n=150$ rzutów. Wyniki:

i	1	2	3	4	5	6
N_i	15	27	36	17	26	29

$$H_0: (N_1, N_2, N_3, N_4, N_5, N_6)$$

$$\sim \text{Mult}(150, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$$

$$H_1: \neg H_0$$

$$\chi^2 = \frac{(15-25)^2}{25} + \frac{(27-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(17-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(29-25)^2}{25} = 12,24$$

$$\chi^2_{1-0,05}(5) \approx 11,7$$

→ **odrzucaamy H_0 .**



Test zgodności chi-kwadrat z nieznanym parametrem

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu dyskretnego o k wartościach (ozn. $1, \dots, k$).

H_0 : prawdopodobieństwa w rozkładzie X to

i	1	2	3	...	k
$P(X=i)$	$p_1(\theta)$	$p_2(\theta)$	$p_3(\theta)$	...	$p_k(\theta)$

przy czym θ jest nieznanym parametrem wymiaru d .

$H_1: \neg H_0$ (tzn. rozkład jest jakiś inny)



Test zgodności chi-kwadrat z nieznanym parametrem – postać testu

Statystyki testowe konstruujemy analogicznie jak poprzednio, przy czym wartości oczekiwane wyliczamy w oparciu o estymatory NW parametrów θ . Zmienia się tylko liczba stopni swobody w rozkładzie:

Tw. Przy prawdziwej H_0 rozkład statystyki χ^2 zmierza do rozkładu χ^2 o $k-d-1$ stopniach swobody $\chi^2(k-d-1)$ przy $n \rightarrow \infty$



Test zgodności chi-kwadrat – uciąglenie

Lepsze są testy Kołmogorowa, ale można również stosować test chi-kwadrat

Model: $X_1, X_2, \dots, X_n$ są próbą IID z rozkładu ciągłego.

H_0 : Rozkład zadany jest dystrybuantą F

H_1 : $\neg H_0$ (tzn. rozkład jest jakiś inny)

Wystarczy rozbić zbiór wartości zmiennej na rozłączne przedziały i zliczać obserwacje, które wpadły do poszczególnych przedziałów.

P-stwa oczekiwane są znane (wynikają z F).

Dalej: test chi-kwadrat



Test zgodności chi-kwadrat – uwagi praktyczne

- ☐ Test powinien być stosowany dla dużych prób
- ☐ Liczebności klas (oczekiwane) nie mogą być za małe (<5). Jeśli za małe, należy pogrupować obserwacje
- ☐ Przedziały klas w wersji „ciągłej” ustalane dowolnie, ale warto zadbać o równomierne rozłożenie p-stw teoretycznych.



Test niezależności chi-kwadrat

Model: $(X_1, Y_1), \dots, (X_n, Y_n)$ są próbą IID z rozkładu dwuwymiarowego o $r \times s$ wartościami (ozn. zbiorem $\{1, \dots, r\} \times \{1, \dots, s\}$).

Niech rozkład teoretyczny zadany będzie przez

$$p_{ij} = P(X = i, Y = j) \quad i = 1, \dots, r \quad j = 1, \dots, s$$

Oznaczmy

$$p_{i\cdot} = \sum_{j=1}^s p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^r p_{ij}$$

Interesuje nas hipoteza o niezależności X i Y :

$$H_0: p_{ij} = p_{i\cdot} * p_{\cdot j} \quad i = 1, \dots, r, \quad j = 1, \dots, s$$

$$H_1: \neg H_0$$

Test niezależności chi-kwadrat – cd.

Rozkład empiryczny opisany jest dwuwymiarową tabelką (tzw. **tablica kontyngencji**)

$i \setminus j$	1	2	...	s	$N_{i\bullet}$
1	N_{11}	N_{12}		N_{1s}	$N_{1\bullet}$
2	N_{21}	N_{22}		N_{2s}	$N_{2\bullet}$
...					
r	N_{r1}	N_{r2}		N_{rs}	$N_{r\bullet}$
$N_{\bullet j}$	$N_{\bullet 1}$	$N_{\bullet 2}$		$N_{\bullet s}$	n



Test niezależności chi-kwadrat – postępowanie

Szczególny przypadek testu zgodności z $(r-1) + (s-1)$ parametrami do wyestymowania:

Statystyka testowa:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - N_{i\cdot}N_{\cdot j}/n)^2}{N_{i\cdot}N_{\cdot j}/n}$$

ma rozkład chi-kwadrat z $(r-1)(s-1)$ stopniami swobody (przy prawdziwej H_0)



Test niezależności chi-kwadrat – przykład

Badamy zależność gustów muzycznych i poglądów politycznych, na poziomie $\alpha = 0,05$

	Popieram X	Nie popieram X	Razem
Słucham disco-polo	25	10	35
Słucham rocka	20	20	40
Słucham muzyki klasycznej	15	10	25
Razem	60	40	100

$$\chi^2 = \frac{(25 - 60 * 35 / 100)^2}{60 * 35 / 100} + \frac{(20 - 60 * 40 / 100)^2}{60 * 40 / 100} + \frac{(15 - 60 * 25 / 100)^2}{60 * 25 / 100} + \frac{(10 - 40 * 35 / 100)^2}{40 * 35 / 100} + \frac{(20 - 40 * 40 / 100)^2}{40 * 40 / 100} + \frac{(10 - 40 * 25 / 100)^2}{40 * 25 / 100} \approx 3,57$$

$$\chi^2_{1-0,05}((2-1)(3-1)) = \chi^2_{0,95}(2) \approx 5,99$$

→ **odrzucaamy H_0 .** Źródło: W. Niemiro



Randomizacja testu

Czasem może nie być testu o poziomie istotności równym dokładnie α (np. dla zmiennych o rozkładach dyskretnych).

Wówczas rozwiązaniem jest *randomizacja*.

(TJNM, o ile jest, musi być zrandomizowany).

np. liczba orłów w 8 rzutach, $H_0 : p = 1/2$, $H_1 : p < 1/2$, $\alpha=0,05$:

xi	0	1	2	3	4	5	6	7	8
pi	0,004	0,03	0,11	0,22	0,27	0,22	0,11	0,03	0,004
suma pi	0,004	0,04	0,15	0,36	0,64	0,86	0,97	0,996	1,000

$X \leq 1$ odrzucamy, $X > 2$ OK, $X = 2$: $p = 1/11$ odrzucamy



WARSAW UNIVERSITY
Faculty of Economic Sciences