Jednokierunkowa analiza wariancji

dr Mariusz Grządziel

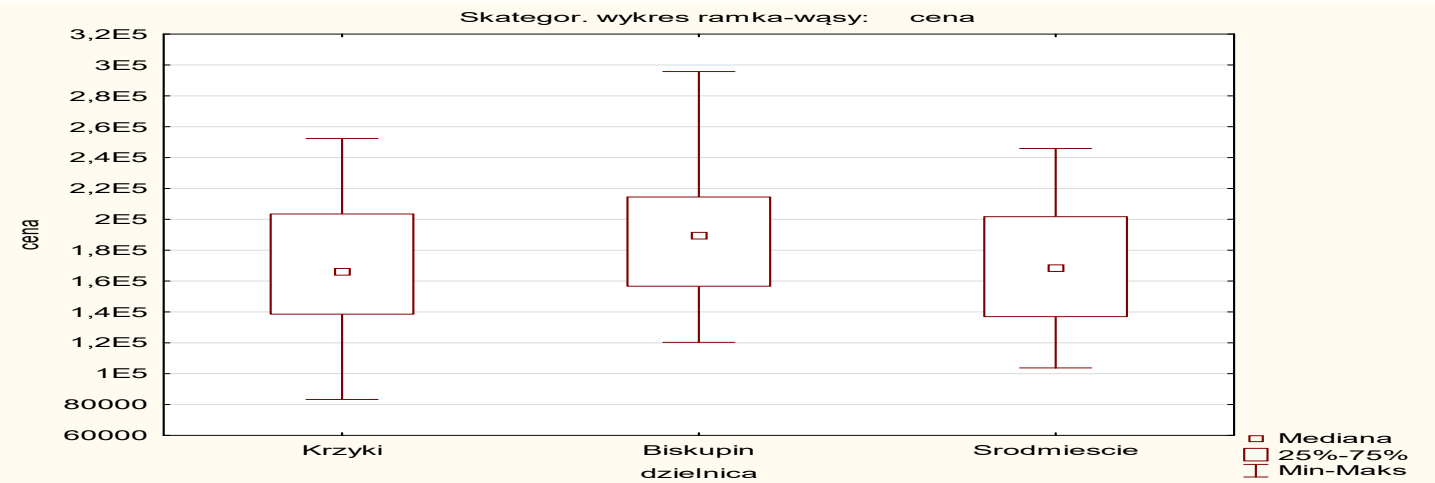
Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

r. akad. 2022/2023

Przykład

Rozważamy dane dotyczące cen mieszkań we Wrocławiu z pakietu Przewodnik (autorstwa P. Biecka) w środowisku R. Dane są dostępne na stronie naszego kursu (w pliku mieszkania.xls)
Jesteśmy zainteresowani porównaniem cen mieszkań w dzielnicach: Biskupin, Śródmieście i Krzyki.

Wykres ramkowy



Formalizacja problemu

Zakładamy, że cecha X_i w i-tej populacji ma rozkład $N(\mu_i, \sigma_i)$ dla $i = 1, \ldots, k$. Niech $X_{i1}, \ldots, X_{in_i}$ będzie próbą z cechy X_i. Jesteśmy zainteresowani weryfikacją hipotezy

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

przeciwko hipotezie H_1, która głosi, że istnieją l i m takie, że $\mu_l \neq \mu_m$.
Przyjmujemy założenie o jednorodności wariancji

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

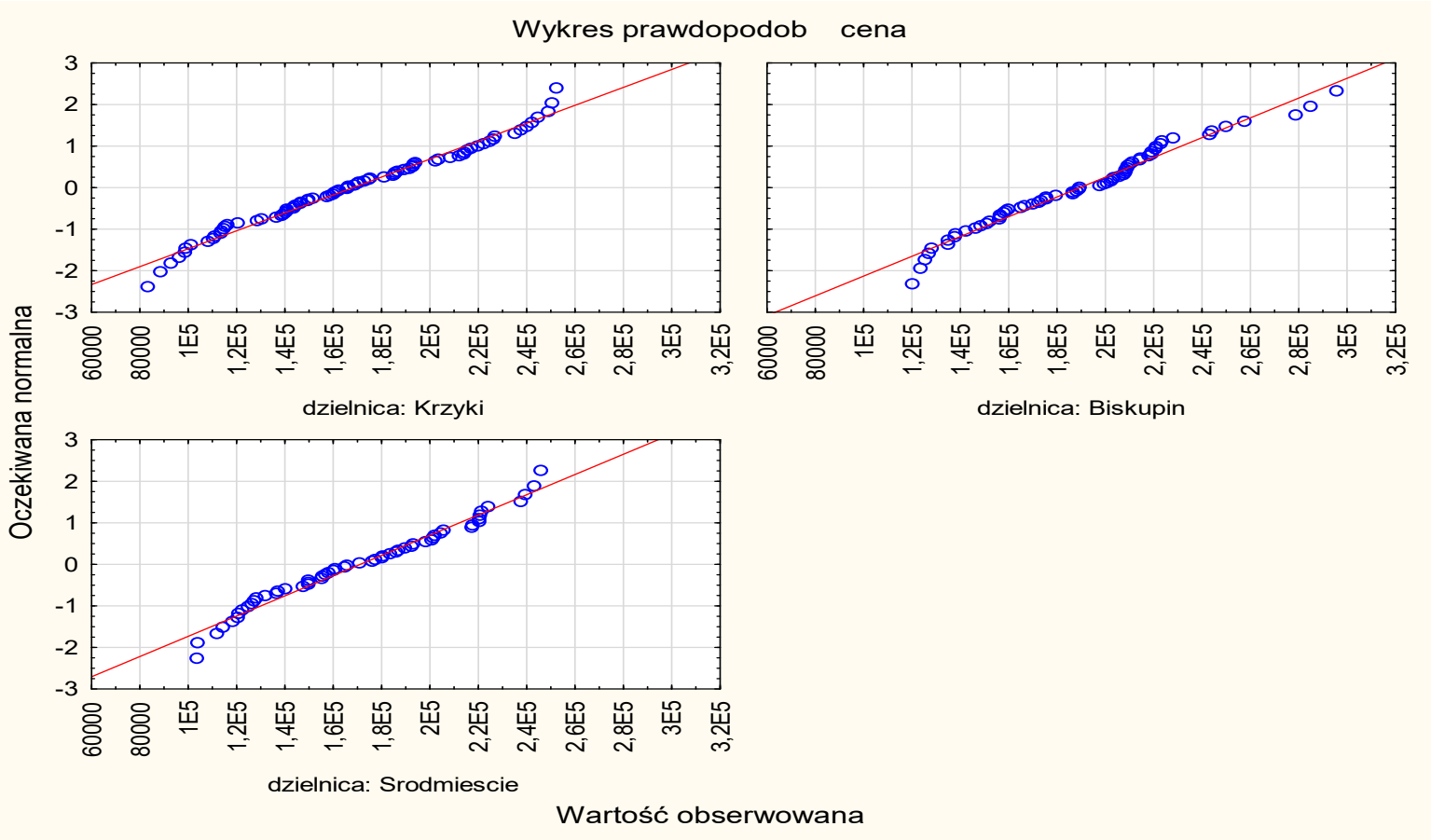
W naszym przykładzie $k = 3$, $n_1 = 79$ (Krzyki), $n_2 = 65$ (Biskupin), $n_3 = 56$ (Śródmieście).

Statystyka/Statystyki Podstawowe/Przekroje, proste ANOVA/Podsum.: tabela statystyk

Tabela przekrojów statystyk opisowych (m) N=200 (Zmienne zależne nie zawierają BD)			
dzielnica	cena Średnie	cena Ważnych	cena Odch.std
Krzyki	168173,0	79	45142,47
Biskupin	189494,0	65	40647,68
Srodmiescie	171143,5	56	39682,43
Ogół	175934,0	200	43078,65

Założenie normalności

Statystyka/Statystyki Podstawowe/Przekroje, proste ANOVA/Testy ANOVA/Skateg. wykres normalności



Założenie normalności

Statystyki/ Statystyki podstawowe/Tabele liczności/Normalność

Zmienna	dzielnica=Biskupin Testy normalności (m)				
	N	maks D	Lillief. p	W	p
cena	65	0,077836	p > ,20	0,968331	0,094168

Założenie normalności

Zmienna	dzielnica=Krzyki Testy normalności (m) N	maks D	Lillief. p	W	p
cena	79	0,064585	p > ,20	0,972764	0,088937

Założenie normalności

Zmienna	dzielnica=Srodmiescie Testy normalności (m) N	maks D	Lillief. p	W	p
cena	56	0,075133	p > ,20	0,963901	0,091940

Założenia — jednorodność wariancji

Statystyka/Statystyki Podstawowe/Przekroje, proste ANOVA/Testy ANOVA/Test Levene'a

Zmienna	Test Levene'a jednorodności wariancji (m) Zaznaczone efekty są istotne z p < ,05000						
	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F
cena	774005792	2	387002896	1,047479E+11	197	531715365	0,727838

Zmienna	Test Levene'a
	p
cena	0,484246

Weryfikacja hipotezy o średnich

Weryfikujemy hipotezę

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

przeciwko hipotezie H_1, która głosi, że istnieją l i m takie, że $\mu_l \neq \mu_m$.
Niech $N = \sum_{i=1}^{k} n_i$; zdefiniujmy:

$$varA = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \quad varE = \sum_{i=1}^{k}\sum_{i=1}^{n_i}(X_{ij} - \bar{X}_i)^2, \quad varT = \sum_{i=1}^{k}\sum_{i=1}^{n_i}(X_{ij} - \bar{X})^2,$$

gdzie

$$\bar{X}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij}, \quad \bar{X} = \frac{1}{N}\sum_{i=1}^{k}\sum_{i=1}^{n_i} X_{ij}.$$

Fakt. $varT = varA + varE$; stąd: Analiza Wariancji.

Statystyka testowa F zdefiniowana wzorem $F = S_a^2/S_e^2$, gdzie

$$S_a^2 = \frac{varA}{k-1}, \quad S_e^2 = \frac{varE}{N-k}$$

ma rozkład F z $(k-1, N-k)$ stopniami swobody. Hipotezę odrzucamy H_0 (i przyjmujemy hipotezę H_1), jeżeli $F > F_{1-\alpha;k-1,N-k}$.

Zmienna	Analiza wariancji (m) Zaznaczone efekty są istotne z p < ,05000					
	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd
cena	*1,799538E+10*	*2*	*8,997692E+09*	*3,513029E+11*	*197*	*1,783263E+09*

Zmienna	Analiza wariancji (m) Zaznaczone efekty są	
	F	p
cena	*5,045633*	*0,007294*

Porównania wielokrotne

W przypadku odrzucenia hipotezy o równości średnich można dla każdej pary (i, j), $i < j$, zweryfikować $H_0 : \mu_i = \mu_j$ przeciwko $H_1 : \mu_i \neq \mu_j$: wykorzystując:

- test LSD (ang. Least Significant Differences — Najmmiejszych Istotnych Różnic, NIR): statystyka testowa

$$t = \frac{\bar{X}_i - \bar{X}_i}{\sqrt{S_e^2}} \sqrt{\frac{n_i n_j}{n_i + n_j}}$$

  przy prawdziwości H_0 ma rozkład t_{n-k} (t-Studenta) z $n - k$ stopniami swobody. Odrzucamy H_0, jeżeli $|t| > t_{1-\alpha/2,n-k}$; dla każdego α otrzymujemy poziom istotności α. Nie można tego powiedzieć o wszystkich parach rozważanych łącznie.
- test HSD (ang. Honest Significant Difference — Rzetelnych Istotnych Różnic, RIR) Tukeya: modyfikacja testu LSD zapewniająca kontrolę błędu dla wszystkich porównań rozważanych łącznie.

dzielnica	Test NIR; Zmienna: cena (m) Zaznaczone różnice są istotne z p < ,05000 {1} M=1682E2	{2} M=1895E2	{3} M=1711E2
Krzyki {1}		*0,002908*	0,687621
Biskupin {2}	*0,002908*		*0,018103*
Srodmiescie {3}	0,687621	*0,018103*	

dzielnica	Test RIR Tukeya; zmienna: cena (m) Zaznaczone różnice są istotne z p < ,05000		
	{1} M=1682E2	{2} M=1895E2	{3} M=1711E2
Krzyki {1}		*0,007255*	0,914511
Biskupin {2}	*0,007255*		*0,045189*
Srodmiescie {3}	0,914511	*0,045189*	

Polecana literatura

T. Górecki, Podstawy statystyki z przykładami w R, BTC, Legionowo 2011, rozdz. 6.3.12.

R. Kala, Statystyka dla przyrodników, Wydawnictwo Uniwersytetu Przyrodniczego w Poznaniu, 2009, rozdz. C1, C2, C4.