Programowanie narzędzi analitycznych – Z03

1 Wykresy

Zadanie 1
Za pomocą polecenia `curve` narysować wykresy funkcji określonych poniższymi wzorami:

a) $y = 2x - 4$

b) $y = x^3 - 2x^2 + x - 6$ dla $x \in [0, 1]$

c) $y = \sin\left(\frac{10\pi}{3}x\right)$ dla $x \in [0, 2\pi]$

Zadanie 2
Za pomocą polecenia `read.csv` wczytać dane z pliku `EngScoResults.csv`. Sporządzić wykresy rozproszenia dla:

a) liczby zdobytych goli przez Anglików,

b) liczby zdobytych goli przez Szkotów w meczach w Szkocji.

2 QQplot

- In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
- Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

Źródło: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot (link)

Zadanie 3
Sporządzić samodzielnie wykres kwantylowy dla zmiennej zapisanej w pliku PNA_Z03.csv zakładając rozkład normalny zmiennej.

Zadanie 4
Wygenerować zmienną `tSt` zawierającą 1000 obserwacji z rozkładu t-Studenta z 3 stopniami swobody. Sporządzić wykres kwantylowy zmiennej z kwantylami z rozkładu

a) normalnego

b) t-Studenta z 3 stopniami swobody

Zadanie 5
Wygenerować zmienną `wyk` zawierającą 10000 obserwacji z rozkładu wykładniczego z parametrem $\lambda = 1$. Sporządzić wykres kwantylowy zmiennej z kwantylami teoretycznymi z rozkładu Gamma(1,1).

3 Funkcje w R

Zadanie 6
Napisać funkcję *silnia*, która dla podanej liczby całkowitej zwróci wartość jej silni.

Zadanie 7
Napisać funkcję *ProcentZer*, która dla podanej macierzy zawierającej 0 i 1 zwróci udział liczby zer w liczbie elementów macierzy.

Zadanie 8
Napisać funkcję *PoleProstokata*, która dla podanych długości boków wylicza pole prostokąta. Jeżeli długość drugiego boku nie będzie podana, to domyślnie ma być przyjmowana wartość

a) 10,

b) długości pierwszego boku.

4 Jarque-Bera test

In statistics, the Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test is named after Carlos Jarque and Anil K. Bera. The test statistic JB is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right) \tag{1}$$

where n is the number of observations (or degrees of freedom in general); S is the sample skewness, C is the sample kurtosis, and k is the number of regressors:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}, \tag{2}$$

$$C = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{2}}, \tag{3}$$

where $\hat{\mu}_3$ and $\hat{\mu}_4$ are the estimates of third and fourth central moments, respectively, $\bar{x}$ is the sample mean, and $\hat{\sigma}^2$ is the estimate of the second central moment, the variance.

If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom, so the statistic can be used to test the hypothesis that the data are from a normal distribution.

Źródło: https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

Zadanie 9
Napisać program/procedurę realizującą test Jarque-Bera. Program powinien zwracać wartość statystyki testowej, p-value oraz zmienną zerojedynkową z wynikiem testu (odrzucenie/nieodrzucenie H_0).

Zadanie 10
Wykorzystać program z zadania 4 do sprawdzenia normalności rozkładu zmiennej z zadania 3.

Zadanie 11
Przeprowadzić eksperyment składający się z 10 tysięcy powtórzeń wygenerowania zmiennej o 100 obserwacjach z rozkładu t-Studenta z 12 stopniami swobody i przeprowadzeniu testu JB. W każdym powtórzeniu zapisujemy wynik (statystykę testową lub wynik zerojedynkowy). Sprawdzić w jakim procencie przypadków test wskazuje prawidłowy wynik.

Zadanie 12
Wygenerować zmienną o 30 obserwacjach z rozkładu normalnego ze średnią 4 i odchyleniem standardowym 2. Sprawdzić normalność rozkładu za pomocą testu JB.

5 Test Durbina-Watsona

Najpopularniejszym testem służącym do weryfikacji hipotezy o braku autokorelacji czynników losowych jest test Durbina-Watsona. Hipotezy zerową i alternatywną w tym teście formułujemy w sposób następujący:

$$\begin{aligned} H_0 &: \quad Cov(\varepsilon_t, \varepsilon_{t-1}) = 0 \\ H_0 &: \quad Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0 \qquad \text{dla } t = 1, \dots, T. \end{aligned}$$

Hipoteza zerowa w teście Durbina-Watsona mówi o braku autokorelacji pierwszego rzędu, czyli braku korelacji między ε_t a ε_{t-1}. Ogólniej o autokorelacji rzędu s mówimy, gdy występuje autokorelacja między ε_t a ε_{t-s}, gdzie $s \geq 1$.

Statystyka DW jest standardowo umieszczana na wydrukach z wynikami pochodzącymi z pakietów ekonometrycznych. Statystykę DW liczymy ze wzoru:

$$\begin{aligned} DW &= \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=2}^{T} e_t^2} = \frac{2\sum_{t=1}^{T} e_t^2 - 2\sum_{t=1}^{T} e_t e_{t-1} - e_1^2 - e_T^2 + 2e_1 e_0}{\sum_{t=1}^{T} e_t^2} \\ &= 2(1 - \hat{\rho}_{\varepsilon_t, \varepsilon_{t-1}}) - \frac{e_1^2 + e_T^2 - 2e_1 e0}{\sum_{t=1}^{T} e_t^2}, \end{aligned}$$

gdzie e_t są resztami z regresji, a $\hat{\rho}$ jest współczynnikiem korelacji empirycznej między e_t i e_{t-1}:

$$\hat{\rho}_{e_t, e_{t-1}} = \frac{\sum_{t=1}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2}.$$

Wnioskowanie statystyczne na podstawie wyliczonej statystyki DW przebiega następująco:

1. Jeśli $DW < 2$:
   a) $DW < d_L$: odrzucamy H_0 i przyjmujemy hipotezę o dodatniej autokorelacji;
   b) $d_L < DW < d_U$: brak konkluzji;
   c) $DW > d_U$: nie ma podstaw do odrzucenia H_0 o braku autokorelacji;
2. Jeśli $DW > 2$:
   a) $DW > 4 - d_L$: odrzucamy H_0 i przyjmujemy hipotezę o ujemnej autokorelacji;
   b) $4 - d_U < DW < 4 - d_L$: brak konkluzji;
   c) $DW < 4 - d_U$: nie ma podstaw do odrzucenia H_0 o braku autokorelacji;

Zadanie 13
Napisać program/procedurę realizującą test Durbina-Watsona. Program powinien zwracać wartość statystyki testowej.

6 Test Ljunga-Boxa

Model powinien się charakteryzować oszacowaniami ε_t (resztami e_t), które są nieskorelowane w czasie. Testem diagnostycznym, który bada hipotezę o nieskorelowaniu ε_t jest test Ljunga-Boxa postaci:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k} \to \chi_m^2,$$

gdzie

$$\hat{\rho}_k = \frac{(T-k)\sum_{t=k+1}^{T}(y_t-\bar{y})(y_{t-k}-\bar{y})}{T^{-1}\sum_{t=1}^{T}(y_t-\bar{y})^2}$$

jest empirycznym współczynnikiem korelacji rzędu k a m jest liczbą współczynników korelacji uwzględnionych w trakcie liczenia statystyki. Na podstawie Skryptu do ekonometrii, J.Mycielskiego, str. 204-205.

$H_0 : \rho_1 = \rho_2 = \ldots = \rho_m = 0$

Test Q Ljunga-Pierce'a ma bardzo podobną statystykę testową:

$$Q_{BP} = n\sum_{k=1}^{h}\hat{\rho}_k^2 \tag{4}$$

oraz wykorzystuje ten sam zbiór krytyczny co test Ljunga-Boxa. Za pomocą symulacji można pokazać, że statystyka Ljunga-Boxa jest lepsza dla wszystkich liczności próbek, również małych. Na podstawie https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test.

Zadanie 14
Napisać program/procedurę realizującą test Ljunga-Boxa. Program powinien zwracać wartość statystyki testowej, p-value oraz zmienną zerojedynkową z wynikiem testu (odrzucenie/nieodrzucenie H_0).

7 Test chi-kwadrat niezależności*

(X, Y) - dwuwymiarowa zmienna losowa o rozkładzie dyskretnym, tzn. $(X, Y) \in \{1, 2, \ldots, r\} \times \{1, 2, \ldots, s\}$;

Niech

$$p_{i,j} = P(X = i \wedge Y = j)$$

$$p_{i\bullet} = P(X = i) = \sum_{j=1}^{s} p_{i,j} \qquad p_{\bullet j} = P(Y = j) = \sum_{i=1}^{r} p_{i,j}$$

$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ próba losowa

$$N_{i,j} = \sum \mathbb{1}(X_l = i \wedge Y_l = j)$$

$$N_{i\bullet} = \sum_{j=1}^{s} N_{i,j} \text{ i } N_{\bullet j} = \sum_{i=1}^{r} N_{i,j}$$

Źródło: Agata Boratyńska, WYKŁADY ZE STATYSTYKI MATEMATYCZNEJ (II rok WNE)

Dane przedstawiamy w tablicy zwanej tablicą kontyngencji.

$x\|y$	1	2	...	s	$N_{i,\bullet}$
1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,s}$	$N_{1,\bullet}$
2	$N_{2,1}$	$N_{2,2}$	...	$N_{2,s}$	$N_{2,\bullet}$
...	...	...	...	...	...
r	$N_{r,1}$	$N_{r,2}$	...	$N_{r,s}$	$N_{r,\bullet}$
$N_{\bullet,j}$	$N_{\bullet,1}$	$N_{\bullet,2}$	...	$N_{\bullet,s}$	N

Hipoteza zerowa: H_0 : X i Y są niezależne.

$$H_0 : p_{i,j} = p_{i,\bullet} * p_{\bullet,j}, \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, s.$$

Nieznanymi parametrami są: $p_{i,\bullet}$ i $p_{\bullet,j}$

Ich estymatory największej wiarogodności to:

$$\hat{p}_{i,\bullet} = \frac{N_{i,\bullet}}{N} \qquad \hat{p}_{\bullet,j} = \frac{N_{\bullet,j}}{N}$$

Estymatory parametrów $p_{i,j}$ są postaci

$$\hat{p}_{i,j} = \hat{p}_{i,\bullet} * \hat{p}_{\bullet,j} = \frac{N_{i,\bullet}}{N} \frac{N_{\bullet,j}}{N}$$

Statystyka testu chi-kwadrat ma postać

$$\chi^2_{test} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{i,j} - \frac{N_{i,\bullet} N_{\bullet,j}}{N}\right)}{\frac{N_{i,\bullet} N_{\bullet,j}}{N}}$$

Jeżeli n dąży do ∞ to rozkład statystyki χ^2_{test} dąży do rozkładu $\chi^2(r-1)(s-1)$.

Hipotezę H_0 odrzucamy gdy $\chi^2_{test} > \chi^2(\alpha, (r-1)(s-1))$

Źródło: Agata Boratyńska, WYKŁADY ZE STATYSTYKI MATEMATYCZNEJ (II rok WNE)

Zadanie 7 (Zadanie 11.2 z [2])
W celu zbadania zależności pomiędzy płcią klientów i ich preferencjami, wylosowano próbę 200 kobiet i mężczyzn i zadano im pytanie: czy uważasz za lepszy produkt firmy A czy B? Wyniki były następujące

Wybrany produkt	kobiety	mężczyźni
wolę A	20	45
wolę B	60	15
nie widzę różnicy	40	20

Zweryfikować hipotezę mówiącą, że preferencje klientów nie zależą od płci, na poziomie istotności 0,10.

Zadanie 8 (Zadanie 11.5 z [2])
Badano związek pomiędzy wykształceniem a zarobkami. Wykształcenie każdej z badanych osób sklasyfikowano jako podstawowe, średnie lub wyższe. Zarobki zostały sklasyfikowane na trzech poziomach. Wyniki przedstawia poniższa tabela.

	podstawowe	średnie	wyższe
< 1000	54	78	128
$1000 - 2000$	75	122	73
> 2000	71	40	49

Zweryfikować hipotezę o niezależności obu cech na poziomie istotności 0,025.

Zadanie 9 (Zadanie 11.6 z [2])
Pewien produkt można wytwarzać trzema metodami produkcji. Wysunięto hipotezę, że wadliwość nie zależy od metody produkcji. Wylosowano niezależnie od metody produkcji próbę 270 sztuk i otrzymano wyniki

Metoda produkcji	jakość dobra	jakość zła
metoda I	40	10
metoda II	80	60
metoda III	60	20

8 Bibliografia

[1] Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M., *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, Tom I i II, Wydawnictwo Naukowe PWN, Warszawa 2010.

[2] Boratyńska A., *Zadania na ćwiczenia ze statystyka matematycznej.*

