Wykład 7: Testowanie zgodności

SPRAWDZANIE NORMALNOŚCI ROZKŁADU

I. Metody graficzne sprawdzania normalności

1. Wykres skrzynkowy. Jeśli nie jest on symetryczny lub widać na nim dużo obserwacji odstających (dla rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$ średnio 7 obserwacji na 1000 znajduje się poza przedziałem
$$(Q_1 - 1.5IQR, Q_3 + 1.5IQR),$$
to uznajemy, że rozkład cechy w próbie znacznie odbiega od rozkładu normalnego. W przeciwnym przypadku, dane mogą, ale nie muszą, pochodzić z rozkładu normalnego i trzeba to sprawdzać innymi metodami.

```
> boxplot(dane)
```

Niestety wykresy skrzynowe są wiarygodne jedynie dla dużej liczności prób (jeśli mamy mało danych, to nie warto na nie patrzeć podczas badania normalności rozkładu).

2. Histogram częstości i jądrowy estymator gęstości.

```
> hist(dane,freq=F)
> lines(density(dane))
```

Pierwsza z powyższych komend rysuje w R histogram częstości (nie liczności); druga nanosi na poprzednio wykonany rysunek jądrowy estymator gęstości.

3. Wykres kwantylowy (wykres normalności) (Q-Q plot: quantile versus quantile plot).
Niech $x_1, x_2, \ldots, x_n$ oznacza realizację próby losowej. Po jej uporządkowaniu (od obserwacji najmniejszej do największej) otrzymujemy tzw. *statystyki porządkowe* z próby, oznaczane $x_{1:n} \leqslant x_{2:n} \leqslant \ldots \leqslant x_{n:n}$. Wykres kwantylowy to zbiór punktów o współrzędnych $(u_{(i-0,5)/n}, x_{i:n})$, gdzie $i = 1, 2, \ldots, n$ zaś $u_{(i-0,5)/n}$ to kwantyl standardowego rozkładu normalnego rzędu $(i - 0, 5)/n$.

Jeśli próba losowa pochodzi z rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$, to wykres kwantylowy jest zbiorem punktów leżących mniej-więcej na prostej $y = \sigma x + \mu$.

```
> qqnorm(dane)
> qqline(dane)
```

Pierwsza z powyższych komend rysuje w R wykres kwantylowy; druga nanosi na ten wykres linię przechodzącą przez kwartyle.

II. Testy normalności

W testach tych weryfikujemy hipotezę

H_0 : rozkład, z którego pochodzi badana próba losowa, jest normalny

przeciwko hipotezie

H_1 : rozkład, z którego pochodzi badana próba losowa, nie jest normalny.

Poniżej wymieniamy podstawowe testy normalności w kolejności od uznawanego za najlepszy do uznawanego za najsłabszy:

1. Test Shapiro-Wilka

   Zaproponowany w 1965 r. jest to dziś uznawany za najlepszy test uniwersalny normalności rozkładu. Konstrukcja tego testu opiera się na wykresie kwantylowym. Dokładniej, wyznacza się linię, która jest możliwie najlepiej dopasowana do punktów tego wykresu (mówiąc precyzyjniej, wyznacza się tzw. *prostą regresji*) i następnie bada się stopień dopasowania tych punktów do owej prostej.

   ```
   > shapiro.test(dane)
   ```

2. Test Andersona-Darlinga
3. Test Craméra-von Misesa
4. Test Lilliefors'a