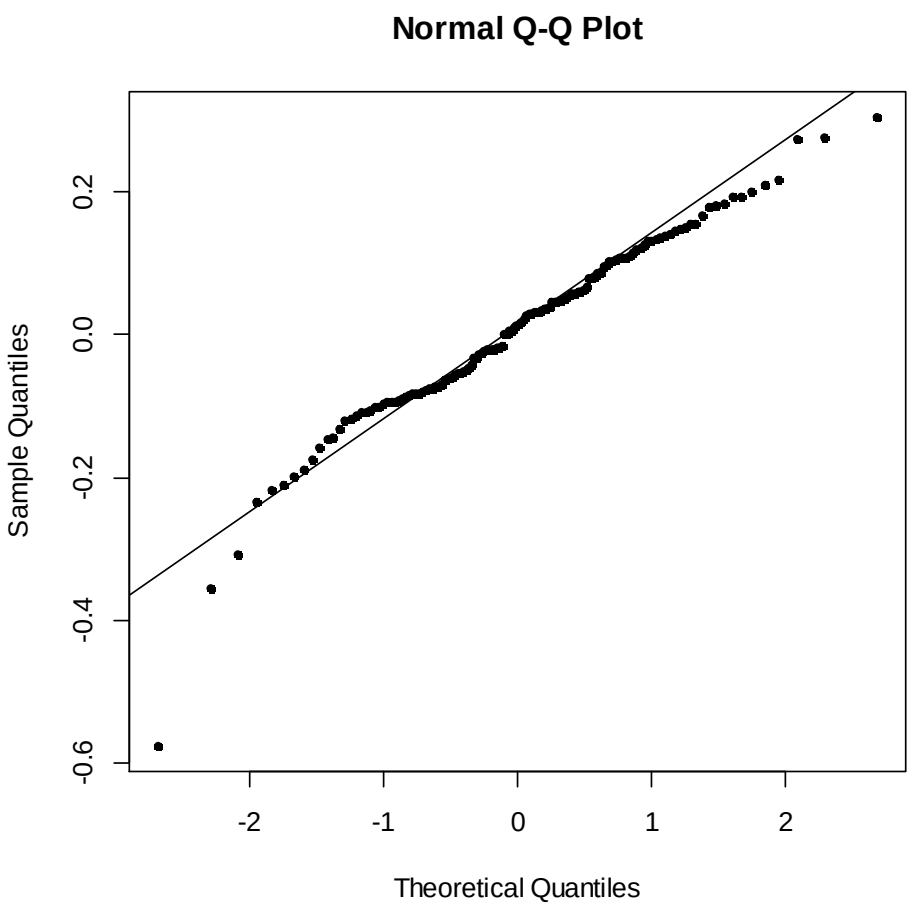
Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

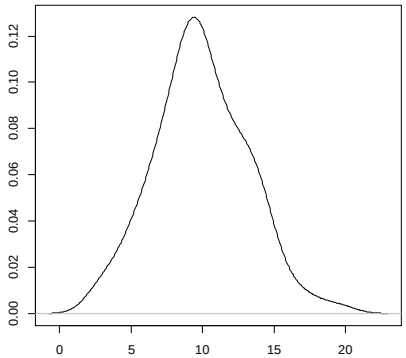
Testowanie zgodności z rozkładem normalnym

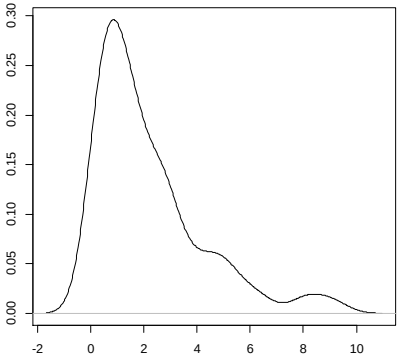
Wykres kwantylowy

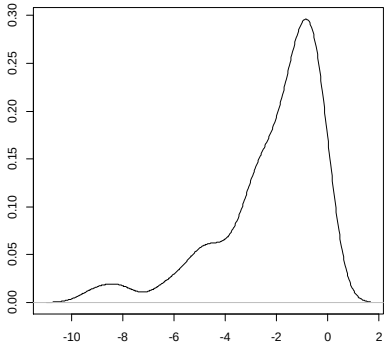


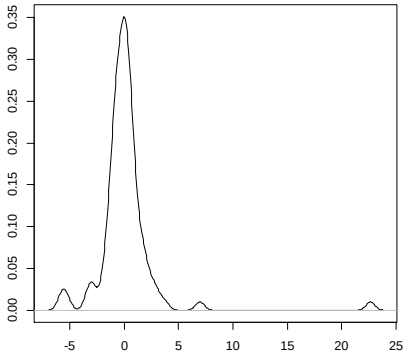
- dane posortowane: $x_1 \leq \cdots \leq x_n$
- wartości x_i są przybliżeniem kwantyla rzędu i/n rozkładu, z którego pochodzą dane,
- porównujemy je z kwantylami $z_{i/n}$ rozkładu normalnego $N(0,1)$
- jeżeli dane pochodzą z rozkładu normalnego, to wykres tworzy linię prostą

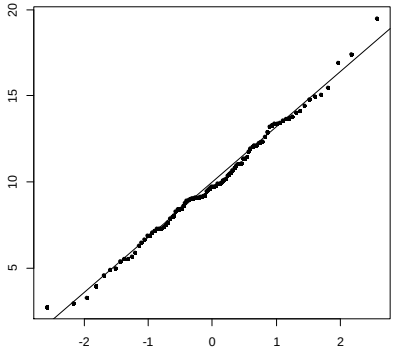
Testowanie zgodności z rozkładem normalnym

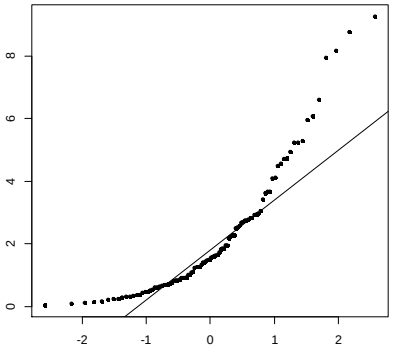


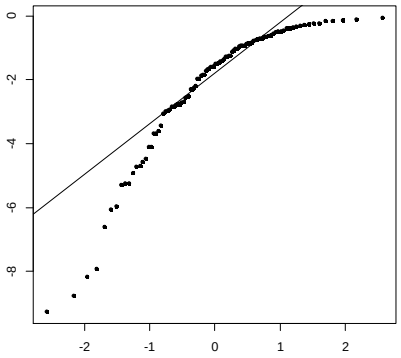


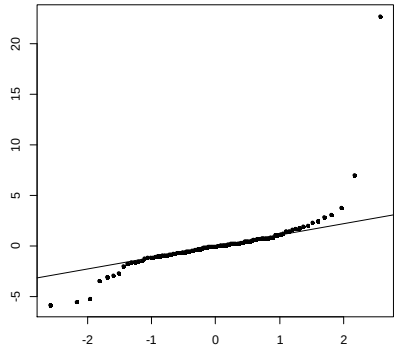












Testowanie zgodności z rozkładem normalnym

Badamy następujące hipotezy:

$$H_0: F = F_0$$

$$H_1: F \neq F_0$$

gdzie F_0 jest dystrybuantą wybranego rozkładu teoretycznego.

F_0 może oznaczać dystrybuantę konkretnego rozkładu (hipoteza prosta) lub rodzinę rozkładów (hipoteza złożona).

W przypadku testów normalności:

- F_0 - dystrybuanta rozkładu normalnego.

Testowanie zgodności z rozkładem normalnym

Test Shapiro-Wilka:

- opiera się na wykresie kwantylowym,
- ma dużą moc (zwłaszcza gdy rozkład jest wyraźnie skośny lub gdy jest symetryczny, ale spłaszczony),
- może być stosowany do małych prób,
- statystyka ma niestandardowy rozkład (zależny od liczności próby).

Test Jarque-Bera:

- opiera się na obserwacji, że rozkład normalny jest symetryczny i ma kurtozę równą 3.

Statystyka ma postać:

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

gdzie:

$A = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$, $K = \frac{\hat{\mu}_4}{\hat{\sigma}^4}$ są estymatorami skośności i kurtozy.

Przy założeniu prawdziwości hipotezy H_0 statystyka JB ma asymptotycznie rozkład χ^2 o dwóch stopniach swobody.

Testy zgodności

Test Kołmogorowa:

- służy do badania zgodności z rozkładem ciągłym,

Dla zaobserwowanych wartości $x_1 \leq \cdots \leq x_n$ definiujemy dystrybuantę empiryczną:

$$F_n(x) = \begin{cases} 0 & x < x_1 \\ k/n & x_k \leq x < x_{k+1} \\ 1 & x \geq x_n \end{cases}$$

Wtedy statystyka testowa ma postać:

$$\lambda = D\sqrt{n},$$

gdzie $D = \sup_x |F_n(x) - F_0(x)|$.

Wartości statystyki λ są stablicowane. Gdy $\lambda > \lambda_\alpha$ to odrzucamy H_0.

Testy zgodności

Test Kołmogorowa:

- hipoteza główna powinna być hipotezą prostą,
- jeżeli hipoteza H_0 jest prawdziwa, to rozkład statystyki λ nie zależy od rozkładu F_0,
- jeżeli hipoteza główna jest hipotezą złożoną, to należy wyestymować parametry rozkładu F_0 należy wyestymować (za pomocą metody największej wiarygodności),
- rozkład statystyki λ jest różny w przypadku, gdy hipoteza główna jest prosta (znamy parametry rozkładu) lub złożona (parametry rozkładu trzeba wyestymować),
- największe różnice $F_n(x) - F_0(x)$ są zwykle w okolicach wartości oczekiwanej.

Test Andersona-Darlinga

- uwzględnia to, że różnice wartości dystrybuant (nawet bardzo różnych rozkładów) są bardzo małe w ogonach,
- jest to wersja testu Cramera – von Misesa.

Statystyka ma postać:

$$A^2 = \int_{-\infty}^{+\infty} \frac{\left(F_n(x) - F_0(x)\right)^2}{F_0(x)\left(1 - F_0(x)\right)} dF_0(x)$$

W praktyce oblicza się ją jako:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln F_0(x_i) + \ln(1 - F_0(x_{n+1-i})\right]$$

Testy zgodności

Test χ^2:

- przeznaczony głównie do badania zgodności z rozkładem skokowym,
- może być stosowany do badania zgodności z rozkładem ciągłym,
- w przypadku rozkładu ciągłego: duża utrata informacji związana z dyskretyzacją danych,
- wartość statystyki (a więc i wyniki testu) zależą od przyjętego podziału na klasy.

Test χ^2

Statystyka ma postać:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

gdzie:

k - liczba klas, na które zostały podzielone dane,

n_i - liczebność empiryczna i-tej klasy,

$\hat{n}_i$ - liczebność teoretyczna i-tej klasy (obliczona na podstawie rozkładu z hipotezy H_0),