Estymatory 1

Joanna Czarnowska[1]

[1]Uniwersytet Gdański
Instytut Matematyki

Dane, a rozkłady prawdopodobieństwa

Przykład 1. Rzucamy n ($n = 30$) razy monetą, otrzymujemy wyniki $x_1, x_2, \ldots, x_n$:

$$01101110111000011000111110111 1$$

Jak na podstawie otrzymanych wyników ocenić, czy dana moneta jest symetryczna, czy też nie?

Dane, a rozkłady prawdopodobieństwa

Przykład 1. Rzucamy n ($n = 30$) razy monetą, otrzymujemy wyniki $x_1, x_2, \ldots, x_n$:

$$01101110111000011000111110111 1$$

Jak na podstawie otrzymanych wyników ocenić, czy dana moneta jest symetryczna, czy też nie?

Budując model dla tego doświadczenia wprowadzimy zmienną losową X, która przyjmuje wartość 1 lub 0 w zależności od tego czy wypadnie orzeł (O), czy reszka (R)

$$X(w) = \begin{cases} 0 & \text{dla } w = R \\ 1 & \text{dla } w = O. \end{cases}$$

Zmienna losowa X ma rozkład zero-jedynkowy. Liczba orłów w podanym przykładzie to $x_1 + \ldots + x_n = 19$.

Dane, a rozkłady prawdopodobieństwa

Przykład 1. Rzucamy n ($n = 30$) razy monetą, otrzymujemy wyniki $x_1, x_2, \ldots, x_n$:

$$01101110111000011000111110111 1$$

Jak na podstawie otrzymanych wyników ocenić, czy dana moneta jest symetryczna, czy też nie?

Budując model dla tego doświadczenia wprowadzimy zmienną losową X, która przyjmuje wartość 1 lub 0 w zależności od tego czy wypadnie orzeł (O), czy reszka (R)

$$X(w) = \begin{cases} 0 & \text{dla } w = R \\ 1 & \text{dla } w = O. \end{cases}$$

Zmienna losowa X ma rozkład zero-jedynkowy. Liczba orłów w podanym przykładzie to $x_1 + \ldots + x_n = 19$.

Oznaczmy prawdopodobieństwo wypadnięcia orła przez $P(X{=}1){=}p$.

Nasz problem sprowadza się zatem do wyznaczenia p. Zauważmy dalej, że

$$E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p.$$

Powstaje więc pytanie, jak mając dane, wyznaczyć wartość oczekiwaną rozkładu?

Dane, a rozkłady prawdopodobieństwa

Otrzymany ciąg wyników $x_1, x_2, \ldots, x_n$ w Przykładzie 1, jest jedną z możliwych realizacji ciągu zmiennych losowych $X_1, X_2, \ldots, X_n$, mających taki sam zero-jedynkowy rozkład, jak zmienna losowa X.

Ciąg zmiennych losowych $X_1, X_2, \ldots, X_n$ będziemy nazywali próbą losową. Próbę losową nazywamy prostą, jeżeli zmienne losowe $X_1, X_2, \ldots, X_n$ mają taki sam rozkład i są niezależne.

Aby z próby wyznaczyć wartości parametrów rokładu, takich jak na przykład wartość oczekiwana, wariancja, skośność – wprowadzimy pojęcie estymatora.

Estymatory wartości oczekiwanej i wariancji

Mamy zmienne losowe $X_1, X_2, \ldots X_n$ niezależne o jednakowym rozkładzie, o wartości oczekiwanej μ i wariancji σ^2. Definiujemy dwie zmienne losowe

1. średnią z próby

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. wariancję z próby

$$\tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2.$$

Średnia z próby i wariancja z próby są przyładami estymatorów, czyli zmiennych losowych $T(X_1, X_2, \ldots, X_n)$, służących do wyliczenia nieznanych parametrów rozkładu.

Dla próby z Przykładu 1 mamy $\overline{X}_{30} \approx 0{,}63$ i $\tilde{S}_n^2 \approx 0{,}23$.

Estymatory wartości oczekiwanej i wariancji

- Czy średnia i wariancja z próby podane na poprzednim slajdzie są „dobrymi" estymatorami wartości oczekiwanej μ i wariancji σ^2?
- I co to znaczy, że estymator jest „dobry"?

Zobacz Przykład 1 i 2 w skrypcie Estymatory-wykład.R

Estymatory nieobciążone

Definicja

Estymator $\Theta_n = T(X_1, \ldots, X_n)$ parametru θ nazywamy nieobciążonym, jeżeli

$$E(\Theta_n) = \theta.$$

Natomiast, jeżeli

$$\lim_{n \to \infty} E(\Theta_n) = \theta,$$

to estymator nazywamy asymptotycznie nieobciążonym.

Estymatory nieobciążone

Średnia z próby $\overline{X_n}$ jest nieobciążonym estymatorem wartości oczekiwanej μ, natomiast wariacja z próby $\tilde{S}_n^2$ jest estymatorem obciążonym rzeczywistej wariancji rozkładu σ^2. Mamy

- $E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$,
- $E(\tilde{S}_n^2) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{n-1}{n}\sigma^2$.

Wariancja $\tilde{S}_n^2$ jest natomiast asymptotycznie nieobciążonym estymatorem wariancji σ^2

$$\lim_{n\to\infty} E(\tilde{S}_n^2) = \lim_{n\to\infty}\left(\frac{n-1}{n}\sigma^2\right) = \sigma^2.$$

Nieobciążonym estymatorem wariancji jest więc

- $S_n^2 = \frac{n}{n-1}\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$.

Zobacz Przykład 3 w skrypcie Estymatory-wyklad.R

Obciążenie estymatora

Niech $\hat{\theta} = T(X_1, \ldots, X_n)$ będzie obciążonym estymatorem nieznanego parametru θ, czyli $E(\hat{\theta}) \neq \theta$.

- Obciążenie estymatora definiujemy jako

$$Bias(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta.$$

Odejmując od danego estymatora obciążenie, otrzymamy estymator nieobciążony

$$\hat{\theta}_{nieob} = \hat{\theta} - Bias(\hat{\theta}).$$

Rozkład estymatora. Rozkład średniej empirycznej

Estymator danego parametru jest zmienną losową, możemy więc badać jego rozkład.

Przykład 2. Załóżmy, że wzrost pewnej populacji dobrze modeluje rozkład normalny $N(175, 5^2)$.

Wygenerujemy z tego rozkładu $N = 100$ próbek liczności odpowiednio $n = 3, 10, 30$ i 100 oraz obliczymy z nich średnie empiryczne, korzystając odpowiednio z estymatorów

$$\overline{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3),$$

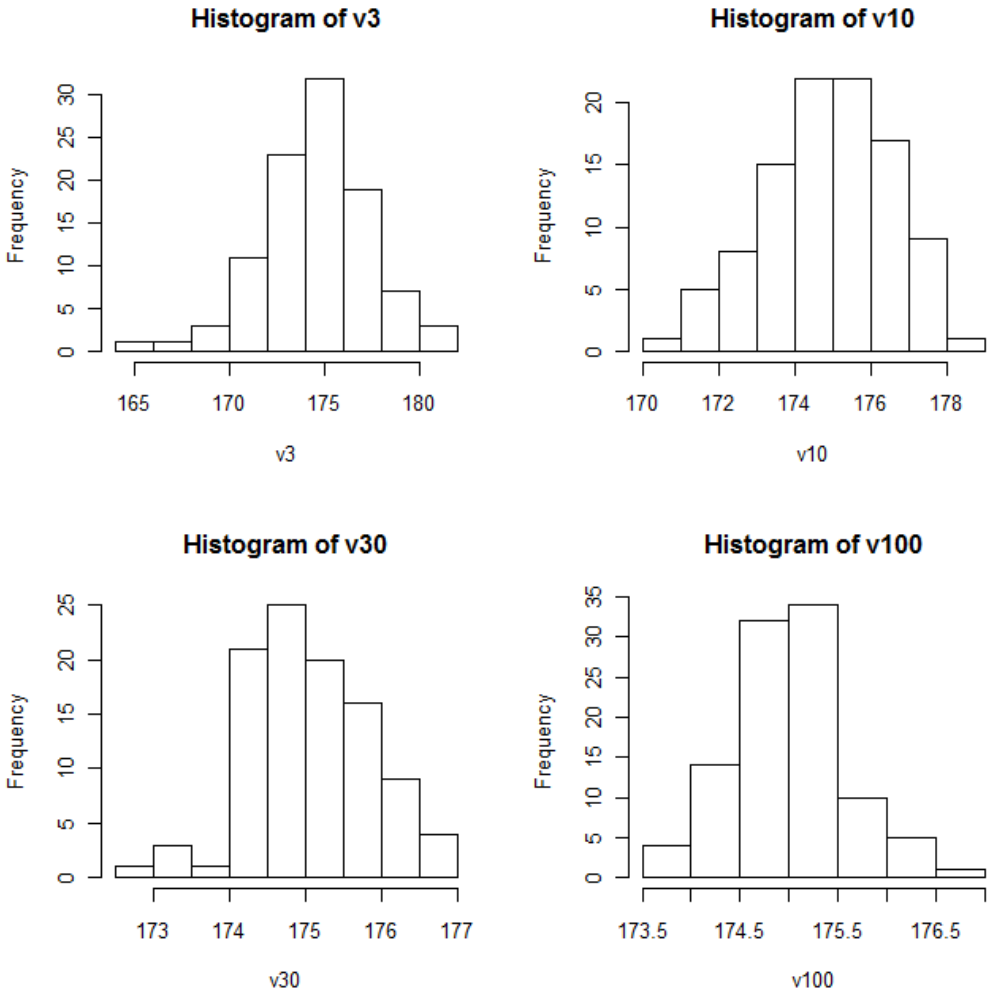
$$\overline{X}_{10} = \frac{1}{10}(X_1 + X_2 + \ldots + X_{10}),$$

$$\overline{X}_{30} = \frac{1}{30}(X_1 + X_2 + \ldots + X_{30}),$$

$$\overline{X}_{100} = \frac{1}{100}(X_1 + X_2 + \ldots + X_{100}).$$

Na kolejnym slajdzie mamy wyniki estymacji.

Rozkład średniej empirycznej



100 próbek 3-element.
100 próbek 10-element.
100 próbek 30-element.
100 próbek 100-element.

Średnie ze 100 średnich:

174.6, 174.8, 175.1, 175.0

Zauważmy, że średnie wyliczane z liczniejszych próbek bardziej skupiają się wokół rzeczywistej średniej (175).

SEM – błąd standardowy średniej

Możemy zadać pytanie, jaki błąd popełniamy, gdy korzystamy ze średniej empirycznej $\overline{X}_n$ do wyznaczenia nieznanej wartości oczekiwanej μ? Zwykle mierzymy go odchyleniem standardowym estymatora.

Łatwo wykazać, że

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Odchylenie standardowe średniej empirycznej $\overline{X}_n$ nazywane jest błędem standardowym średniej (SEM - Standard Error of the Mean)

$$SEM = \frac{\sigma}{\sqrt{n}}$$

i służy do oceny średniego błędu jaki popełniamy, gdy korzystamy ze średniej empirycznej do wyznaczenia nieznanej wartości oczekiwanej μ?

Zauważmy, że σ^2 jest zwykle nieznane – jest estymowane za pomocą estymatora nieobciążonego S_n^2 z póby. Jednocześnie im większa próba, tym oczywiście błąd standardowy średniej jest mniejszy.

Średnia z próby, a rzeczywista wartość.

To właśnie obserwowalismy na histogramach. Im liczniejsza próba z której wyliczaliśmy szukaną średnią, tym wyniki były bliżej skupione 175. Kolejne bowiem odchylenia wynosiły: $\frac{5}{\sqrt{3}}$, $\frac{5}{\sqrt{10}}$, $\frac{5}{\sqrt{30}}$, $\frac{5}{\sqrt{100}}$, czyli 2.9, 1.6, 0.9 i 0.5.

Podsumowując nasz przykład: jeśli szukaną średnią będziemy wyliczali z próbki 3-elementowej, to średni błąd wyniesie $SEM = 2,9$. Jeśli użyjemy do tego próbę 10-elementową, to wyniki będą bardziej skupione wokół szukanej średniej – błąd st. średniej jest 1,6 – jest więc większa szansa, że średnia wyliczona przez nas z próby niewiele różni się od rzeczywistej średniej. Ta nazwijmy wiarygodność wyniku wzrasta jeszcze bardziej przy próbie 30-elementowej – odchylenie jest zaledwie 0,9. No i oczywiście im wiecej tym lepiej, bo przy próbie 100-elementowej mamy odchylenie zaledwie 0,5.

Dystrybuanta empiryczna

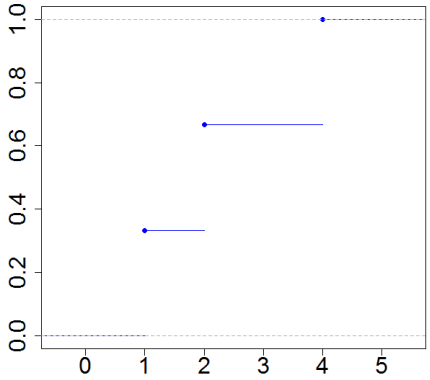
Nieobciążonym estymatorem wartości dystrybuanty F, zmiennej losowej X, w punkcie x jest

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leqslant x\}} \qquad \left(\frac{\text{liczba elementów w próbie} \leqslant x}{n}\right)$$

Funkcję $\hat{F}_n$ nazywamy dystrybuanta empiryczną – jest to funkcja schodkowa.

Przykład 3. Dla uproszczenia obliczeń weźmy próbę złożoną tylko z trzech elementów: 1, 2 i 4. Dystrybuanta empiryczna dana jest wzorem

$$\hat{F}(x) = \begin{cases} 0 & \text{dla } x < 1 \\ 1/3 & \text{dla } 1 \leqslant x < 2 \\ 2/3 & \text{dla } 2 \leqslant x < 4 \\ 1 & \text{dla } x \geqslant 4. \end{cases}$$



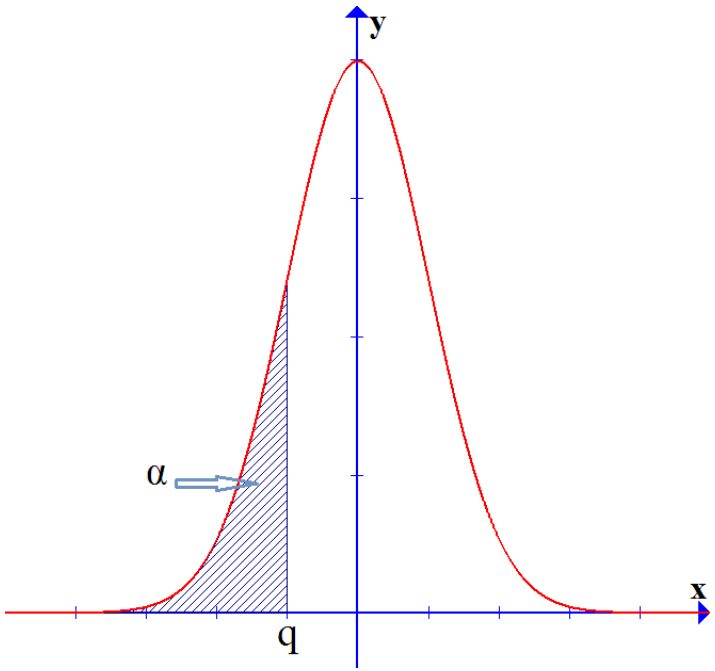
Zobacz Przykład 4 w skrypcie 03Estymatory.R-wyklad

Kwantyle

Kwantyl rzędu $\alpha \in (0,1)$ zmiennej losowej ciągłej X to taka liczba q, dla której prawdopodobieństwo, że zmienna X przyjmuje wartości mniejsze lub równe q jest równe α

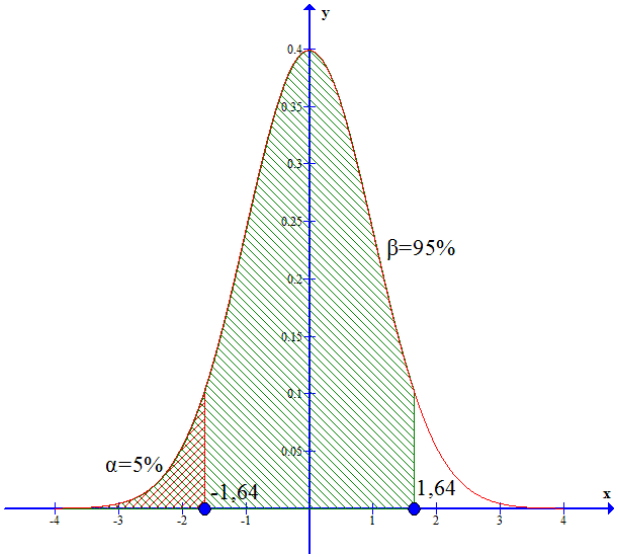
$$P(X \leqslant q) = \alpha.$$

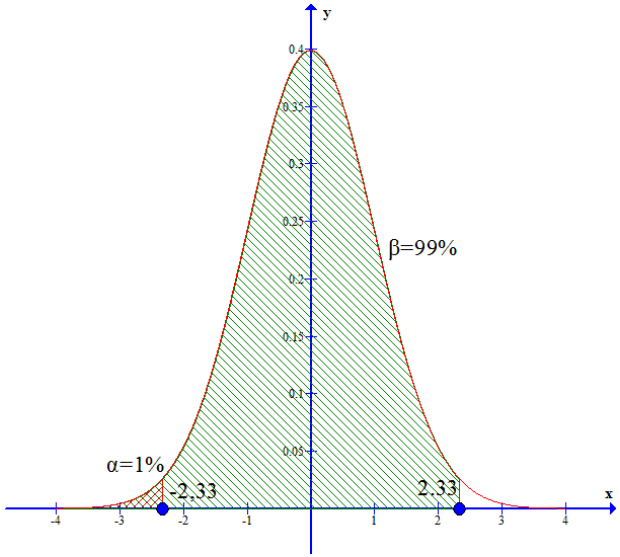
Kwantyl rzędu 50% nazywamy medianą.



Funkcję $q : (0,1) \to \mathrm{R}$, która dowolnej liczbie α z przedziału $(0,1)$, przyporządkowuje kwantyl zmiennej losowej X rzędu α, nazywamy funkcją kwanylową tej zmiennej.

Kwantyle rozkładu normalnego standardowego





- Kwantyl rzędu 95% rozkładu normalnego standardowego to 1,64, natomiast rzędu 99% – 2,33. Prawdopodobieństwo, że zmienna losowa o tym rozkładzie przekroczy wartość 1,64 wynosi więc 5%, natomiast wartośc 2,33 – jedynie 1%.
- Rozkład normalny standardowy jest symetryczny, stąd

$$q_\alpha = -q_{1-\alpha}.$$

Dla przykładu $q_{5\%}=-q_{95\%}=-1,64$, czy $q_{1\%}=-q_{99\%}=-2,33$.

Zobacz Przykład 5 w skrypcie 03Estymatory.R-wyklad

Kwantyle empiryczne

Jest wiele estymatorów kwantyli – omówimy estymator wykorzystywany domyślnie w pakiecie R. Próbę porządkujemy niemalejąco

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

- wartości najmniejszą $x_{(1)}$ w próbie i wartość największą $x_{(n)}$ przyjmuje się jako kwantyle rzędu 0 i 1,
- pozostałe wartości to kwantyle odpowiednio rzędów $\frac{1}{n-1}, \frac{2}{n-1}, \ldots, \frac{n-2}{n-1}$. Czyli

$$x_{(k)} = q_{\frac{k-1}{n-1}}, \quad k = 2, \ldots, n-1.$$

Kwantyle innych rzędów uzyskujemy poprzez interpolację liniową.

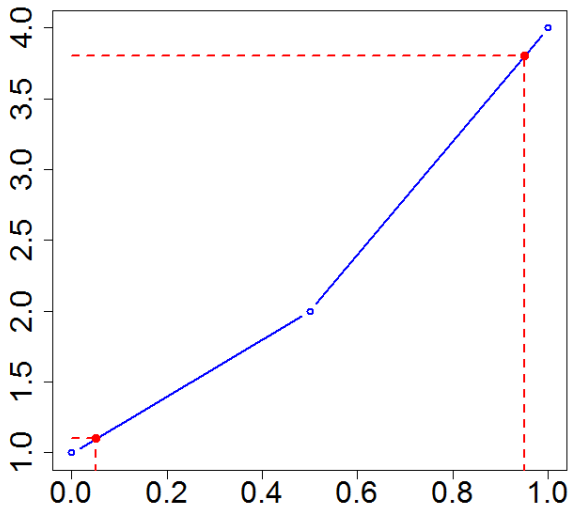
Kwantyle empiryczne

Przykład 4. Weźmy próbę trzyelementową 1, 2, 4. Liczby te są wartością funkcji kwantylowej odpowiednio w punktach 0, $\frac{1}{2}$ i 1.

Kwantyle rzędu 5% i 95% z tej próby

$$\hat{q}_{5\%} = \frac{0,5 - 0,05}{0,5} \cdot 1 + \frac{0,05 - 0}{0,5} \cdot 2 = 1,1$$

$$\hat{q}_{95\%} = \frac{1 - 0.95}{0,5} \cdot 2 + \frac{0,95 - 0,5}{0,5} \cdot 4 = 3,8$$



Inne estymatory zobacz np. pomoc w R (?quantile)

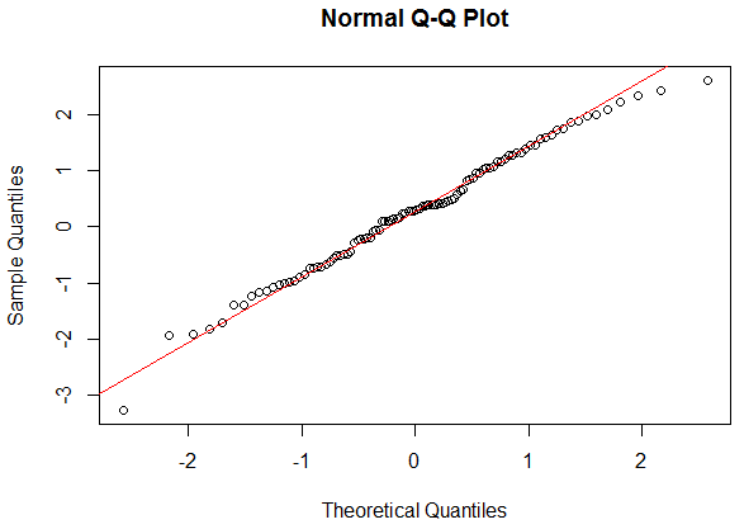
Wykres kwantyl-kwantyl (QQ-plot)

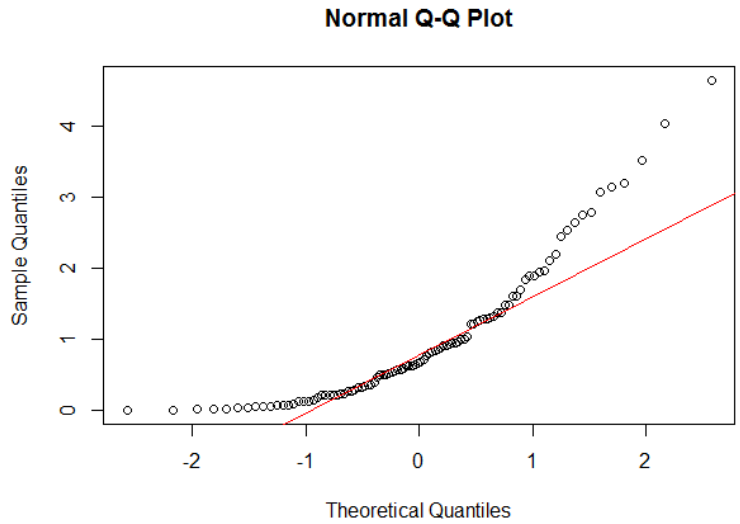
W analizie danych, do oceny na ile dany rozkład teoretyczny dobrze „opisuje" dane, wykorzystujemy wykres *kwantyl-kwantyl*. Jest to wykres

$$\{(q_\alpha, \hat{q}_\alpha) : \ \alpha \in (0,1)\},$$

gdzie q_α to kwantyl teoretyczny rzędu α, a $\hat{q}_\alpha$ – kwantyl empiryczny.
W przypadku dobrego dopasowania punkty położone są na prostej $y = x$ lub blisko tej prostej.

Przykładowo do porównania kwantyli z próby z kwantylami rozkładu normalnego mamy w R funkcję qqnorm().





Zobacz Przykład 6 w skrypcie 03Estymatory.R-wyklad