

## 7. Estymacja parametrów w modelu normalnym(14.04.2008)

### Pojęcie losowej próby prostej

**Definicja 1**  *$n$ -elementową losową próbą prostą nazywamy ciąg  $n$  niezależnych zmiennych losowych o jednakowych rozkładach prawdopodobieństwa.*

Uwaga. Niektórzy autorzy utożsamiają losową próbę prostą  $X_1, X_2, \dots, X_n$  z jej realizacją  $x_1, x_2, \dots, x_n$ ;  $x_1$  jest realizacją  $X_1$  itd.

## Rozkład cen mieszkań

Można oczekiwać, że rozkład cen domów w dzielnicach które nie pretendują do miana „prestizowych” czy „luksusowych” będzie zbliżony do normalnego  $N(\mu, \sigma)$ ; (cena mieszkania jest wtedy sumą niezależnych składników).

Rozważmy dane dotyczące cen mieszkań w dzielnicy A

(dzielnica A jest zdecydowanie mniej prestiżowa niż B) — można założyć, że rozkład cen mieszkań w A będzie zbliżony do normalnego.

## Estymacja średniej $\mu$

Zakładamy, że nasze dane  $x_1, \dots, x_n$  są realizacją losowej próby prostej  $X_1, \dots, X_n$ ;

sensownym oszacowaniem  $\mu$  jest  $\bar{X}$ ;

Dla konkretnej realizacji  $x_1, \dots, x_n$  losowej próby losowej  $X_1, \dots, X_n$  realizacja  $\bar{X}$  jest równa  $\bar{x}$

## Własności zmiennej losowej $\bar{X}$

Zakładamy, że  $X_1, \dots, X_n$  są próbą prostą

$$X_i \sim N(\mu, \sigma), i = 1, 2, \dots, n.$$

$\bar{X} = \frac{1}{n}(X_1, \dots, X_n)$  jest zmienną losową.

Funkcje próby prostej nazywane są *statystykami*.

Z własności wartości oczekiwanej, omawianych na jednym z poprzednich wykładów:

$$E(\bar{X}) = \mu,$$

a więc  $\bar{X}$  jest nieobciążonym estymatorem  $\mu$ .

## Estymacja $\sigma$ i $\sigma^2$

Wariancja z próby  $S^2$  określona przez

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

jest sensownym oszacowaniem  $\sigma^2$ .

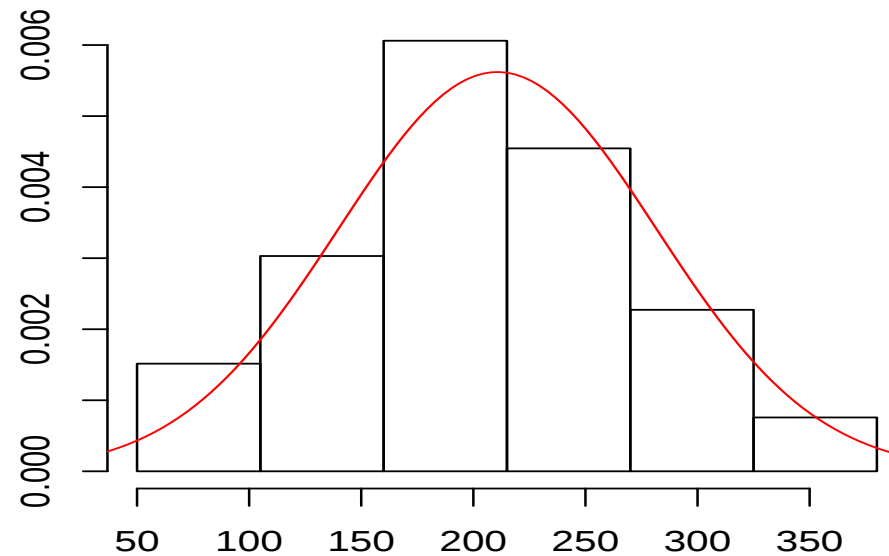
Uwaga: można pokazać, że:

$$E\bar{X} = \mu, \quad E(S^2) = \sigma^2$$

— średnia z próby  $\bar{X}$  jest nieobciążonym (niestronniczym) estymatorem  $\mu$   
wariancja z próby  $S^2$  jest nieobciążonym estymatorem  $\sigma^2$  (por. Koronacki, Mielniczuk, Rozdz. 2.4)

odchylenie standardowe z próby  $S = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$  jest sensownym estymatorem  $\sigma$ , ale nie ma (z reguły) własności nieobciążoności.

## Dane dotyczące cen mieszkań w A- histogram+krzywa normalna



Rysunek 1: Histogram probabilistyczny+ wykres gęstości  $N(\bar{x}, s)$ .

## Weryfikacja założenia o normalności rozkładu populacji

Dane są obserwacje  $x_1, x_2, \dots, x_n$ . Czy można założyć, że  $x_1, x_2, \dots, x_n$  jest realizacją próby prostej z rozkładu normalnego  $N(\mu, \sigma)$  dla pewnych  $\mu$  i  $\sigma$ ?

Metody weryfikacji założenia o normalności:

- Sporządzenie histogramu i porównanie jego kształtu z krzywą dzwonową (wykresem funkcji  $\phi$ .)
- Sporządzenie wykresu ramkowego i odpowiednia jego interpretacja
- Sporządzenie wykresu kwantylowego i odpowiednia jego interpretacja.
- Testy zgodności: Shapiro–Wilka i inne.

Trzy z powyższych metod wykorzystują pojęcie kwantyla.

## Kwantyle rozkładu normalnego— przykład

Założmy, że  $Y$ , wzrost dorosłych mężczyzn w kraju  $A$  ma rozkład normalny  $N(176, 8)$ . Chcemy znaleźć liczbę  $d$  taką, że

$$P(Y < d) = 0,99.$$

Znajomość liczby  $d$  może być przydatna dla inżynierów sporządzających projekty budowlane itd.

Zadanie ogólniejsze: dla liczby  $p \in (0, 1)$  kwantylem rzędu  $p$  rozkładu normalnego  $N(\mu, \sigma)$  nazywamy liczbę  $q_p$  spełniającą równość:

$$P(X < q_p) = \int_{-\infty}^{q_p} \phi_{\mu, \sigma}(x) dx = p,$$

gdzie  $X \sim N(\mu, \sigma)$ . Dla innych rozkładów ciągłych— kwantyle definiujemy analogicznie.

## Obliczanie kwantyli rozkładu normalnego

Korzystając z tablic rozkładu  $N(0, 1)$  : ponieważ

$$P(X < q_p) = P(Z < \frac{q_p - \mu}{\sigma}) = \Phi(\frac{q_p - \mu}{\sigma}),$$

gdzie  $X \sim N(\mu, \sigma)$ ,  $Z \sim N(0, 1)$ , więc zadanie sprowadza się do znalezienia przybliżonej wartości  $c_0$  rozwiązania równania  $\Phi(c) = p$  oraz do wyznaczenia przybliżonej wartości  $q_p$  jako rozwiązania równania:

$$\frac{x - \mu}{\sigma} = c_0.$$

Korzystając z R-a: należy użyć polecenia **qnorm** z odpowiednimi parametrami; w Excelu: należy skorzystać z polecenia ROZKŁAD.NORMALNY.ODW.

## Kwantyle rozkładu normalnego– przykład (c.d.)

$Y$ , wzrost dorosłych mężczyzn w kraju  $A$  ma rozkład normalny  $N(176, 8)$ ; Równanie  $\Phi(c) = 0,99$  ma przybliżone rozwiązanie  $c_0 = 2,33$  (bo  $\Phi(2,32) \approx 0,9898$  i  $\Phi(2,32) \approx 0,9901$ ); przybliżona wartość szukanego kwantyla jest rozwiązaniem równania:

$$\frac{x - 176}{8} = 2,33$$

stąd przybliżona wartość kwantyla rzędu 0,99 rozkładu  $N(176, 8)$  wynosi  $176 + 2,33 \times 8 = 194,64$ .

Korzystając z R-a:

```
> qnorm(0.99, 176, 8)
[1] 194.6108
```

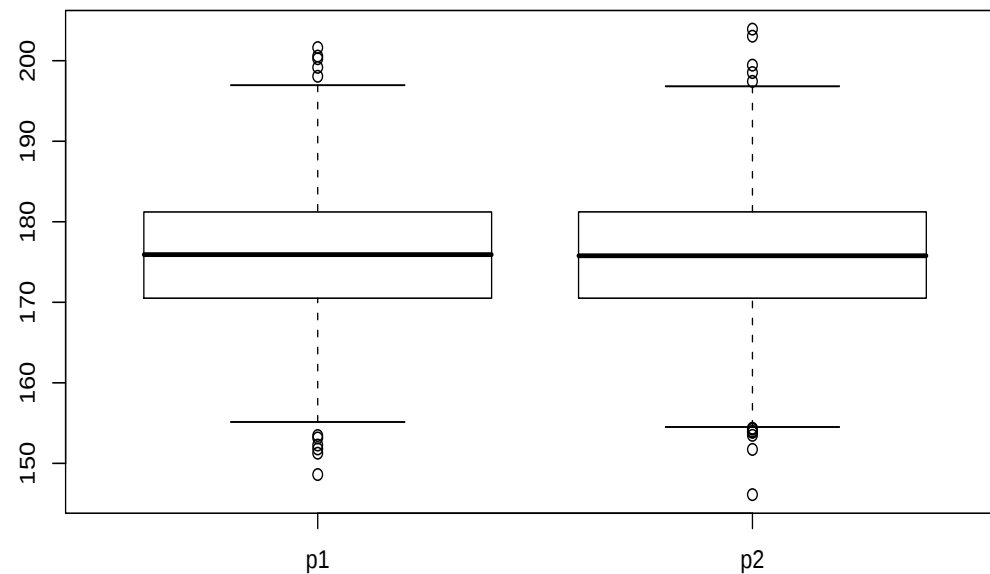
## Wykres ramkowy i założenie normalności rozkładu

Można pokazać że poza przedziałem:

$$[q_{0,25} - 1,5 \times (q_{0,75} - q_{0,25}), q_{0,75} + 1,5 \times (q_{0,75} - q_{0,25})]$$

leży średnio 7 obserwacji na 1000, (jeśli zakładamy, że obserwacje te stanowią realizację próby prostej z rozkładu  $N(\mu, \sigma)$  (por. książkę J. Koronackiego i J. Mielniczuka, par. 1.4.2). Zbyt duża liczba obserwacji odstających wskazuje na to, że założenie normalności rozkładu nie jest spełnione.

Jeśli obserwacje pochodzą z rozkładu normalnego, to odpowiadający im wykres ramkowy będzie w przybliżeniu symetryczny względem odcinka "wewnątrz ramki" (który odpowiada medianie).



Rysunek 2: Wykresy pudełkowe dla dwóch prób o liczności  $n = 1000$ , wygenerowanych przy pomocy generatora liczb pseudolosowych o rozkładzie  $N(176, 8)$

## Wykres kwantylowy

Uporządkowany ciąg obserwacji  $x_1, \dots, x_n$  oznaczamy przez  $x_{1:n}, \dots, x_{n:n}$ . Niech

$$x'_{1:n} = \frac{x_{1:n} - m}{s}, \dots, x'_{n:n} = \frac{x_{n:n} - m}{s},$$

gdzie  $m$  oznacza średnią z próby, a  $s$  odchylenie standardowe z próby. Jeśli założymy, że  $x_1, \dots, x_n$  pochodzą z rozkładu normalnego, to  $x'_{1:n}, x'_{2:n}, \dots$  powinny być sensownymi przybliżeniami kwantyli  $q_{1/n}, q_{2/n}, \dots$  (rzędu  $\frac{1}{n}$ , rzędu  $\frac{2}{n}$  itd.) rozkładu  $N(0, 1)$ ; stąd punkty  $(x'_{1:n}, q_{1/n}), (x'_{2:n}, q_{2/n})$  powinny się układać wokół (pewnej) prostej ; analogicznie punkty:

$$(x_{1:n}, q_{1/n}), (x_{2:n}, q_{2/n}), \dots \tag{1}$$

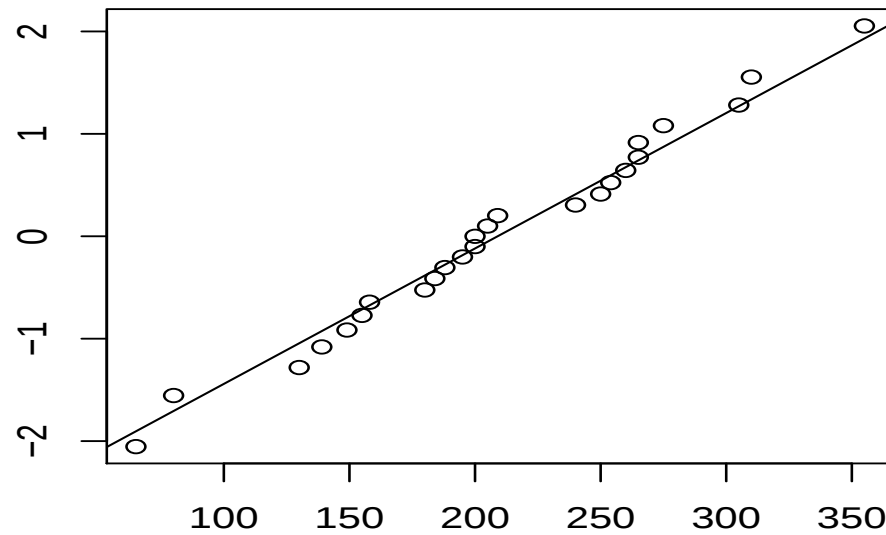
również powinny się układać wokół prostej. Wykres przedstawiający punkty (1) jest nazywany wykresem kwantylowym.

W praktyce zamiast kwantyli rzędu  $i/n$  wyznacza się np. kwantyle rzędu  $i/(n + 1)$ ; w ten sposób unikamy problemu z  $n$ -tym punktem!

Wykres kwantylowy „odbiegający od prostej”: założenie o normalności rozkładu, z którego pochodzą  $x_1, \dots, x_n$ , należy odrzucić.

Szczegóły dotyczące sprawdzania rozkładu kwantylowego można znaleźć w [KM01, par.3.4.2].

## Ceny mieszkań w A



Rysunek 3: Wykres kwantylowy dla danych dotyczących cen mieszkań w A

## **Test Shapiro–Wilka i inne testy zgodności**

Zgodność z rozkładem normalnym można weryfikować za pomocą testu Shapiro-Wilka; w wyniku obliczeń, wykonanych dla otrzymanych obserwacji, otrzymujemy tzw. p-wartość (ang. *p-value*); jeśli p-wartość jest mała, powiedzmy mniejsza niż 0,05, wtedy są podstawy do odrzucenia hipotezy o normalności rozkładu, z którego pochodzą obserwacje.

## Test Shapiro–Wilka: przykład obliczeń

W zmiennej  $y$  zapisane są wartości cen mieszkań w dzielnicy  $A$ .

```
> y
[1] 65 80 139 180 355 158 240 205 265 305 200 155 209 310
[20] 200 184 130 260 250 195
> shapiro.test(y)
```

Shapiro-Wilk normality test

data: y

$W = 0.9849$ ,  $p\text{-value} = 0.962$

$p$ -wartość jest równa 0,962– nie ma podstaw do odrzucenia hipotezy o zgodności z rozkładem normalnym.

Więcej informacji o teście Shapiro-Wilka; [KM01,par. 3.4.2]