Testowanie hipotez w rodzinie rozkładów normalnych— przypadek nieznanego odchylenia standardowego

dr Mariusz Grządziel

Wykład 10; 23 kwietnia 2012

Przykład wprowadzający
Cena metra kwadratowego (w tys. zł) z dla 14 losowo wybranych mieszkań w mieście A:

$$3{,}75; 3{,}89; 5{,}09; 3{,}77; 3{,}53; 2{,}82; 3{,}16;$$
$$2{,}79; 4{,}34; 3{,}61; 4{,}31; 3{,}31; 2{,}50; 3{,}27.$$

W prasie podano informację, że średnia cena metra kwadratowego mieszkań w A wynosi 3800 zł. Czy powyższe dane potwierdzają to stwierdzenie?

Formalizacja problemu
Zakładamy, że dane dotyczące cen metra kw. mieszkań w A są realizacją losowej próby prostej z rozkładu normalnego $N(\mu, \sigma)$. (W wielu praktycznych sytuacjach założenie o normalności rozkładu ceny metra kwadratowego mieszkań— lub domów — należy odrzucić).
Chcemy zweryfikować:

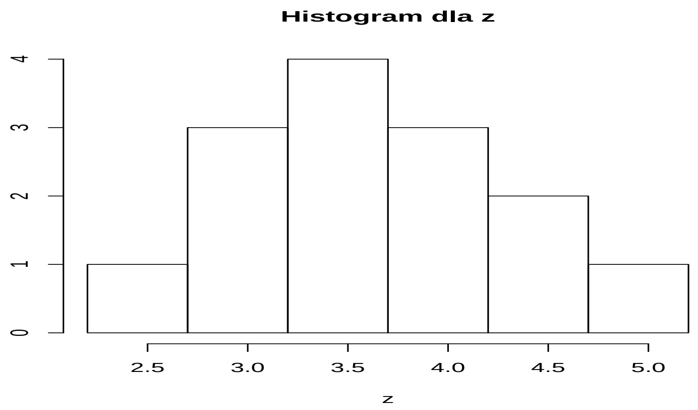
$$H_0 : \mu = \mu_0 \text{ przeciw } H_1 : \mu \neq \mu_0$$

(w naszym przykładzie $\mu_0 = 3{,}8$).

Statystyka testowa i jej rozkład
Zakładamy, że nasze pomiary stanowią realizację próby prostej $X_1, X_2, \ldots, X_n$, gdzie $X_i \sim N(\mu; \sigma)$;
Chcemy „procedurę testową" oprzeć na zmiennej losowej (tzw. *statystyce testowej*)

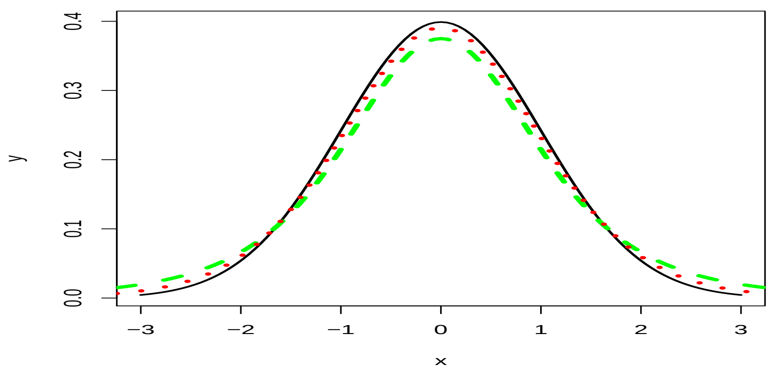
$$T = \frac{\bar{X} - \mu_0}{\mathcal{S}/\sqrt{n}},$$

gdzie $\mathcal{S}$ oznacza odchylenie standardowe z próby $X_1, X_2, \ldots, X_n$.
Postać statystyki testowej T jest podobna do postaci statystyki testowej Z— różnica polega na wstawieniu $\mathcal{S}$ zamiast σ.
Czy można mieć nadzieję, że T będzie miała rozkład $N(0, 1)$ (podobnie jak Z)?



Rysunek 1: Histogram liczebności dla danych dotyczących ceny metra kw. mieszkań w A, wyrażonych w tys. zł; dane są zapisane w zmiennej z

Rozkład statystyki testowej T
Można pokazać, że statystyka testowa T ma rozkład t-Studenta z $n-1$ stopniami swobody; udowodnił to W. Gosset (pseudonim *Student*) na początku XX w. Wykres gęstości rozkładu t-Studenta z $n-1$ stopniami swobody jest „spłaszczony” w stosunku do wykresu gęstości rozkładu normalnego $N(0,1)$ (por. Rys. 1). Analityczne określenie gęstości rozkładu t-Studenta wymaga znajomości funkcji gamma (definiuje się ją za pomocą odpowiedniej całki niekoreślonej por. książkę J. Koronackiego i J. Mielniczuka, str. 201).



Rysunek 2: Wykresy gęstości rozkładów normalnych: normalnego $N(0,1)$ (linia ciągła), t-Studenta z dwoma 4 st. swobody (linia „kreskowana”), t-Studenta z 12 st. swobody (linia „kropkowana”).

Obszar krytyczny

Dla ustalonego poziomu istotności α i hipotezy alternatywnej H_1 obszar krytyczny ma postać:

$$(-\infty, -t_{1-\alpha/2,n-1}] \cup [t_{1-\alpha/2,n-1}, \infty)$$

gdzie $t_{1-\alpha/2,n-1}$ oznacza kwantyl rzędu $1-\alpha/2$ rozkładu t-Studenta z $n-1$ stopniami swobody.

Definicja 1. *Kwantylem rzędu u, $u \in (0,1)$, rozkładu typu ciągłego z funckją gęstości g dodatnią na przedziale I i równą zeru poza tym przedziałem, będziemy nazywali liczbę c_u spełniającą równość:*

$$F(c_u) = u,$$

gdzie F jest dystrybuantą rozkładu (tj. $F(t) = \int_{-\infty}^{t} g(x)dx$).

Przykład— obliczenia
Chcemy zweryfikować hipotezę

$$H_0 : \mu = 3{,}8 \text{ przeciw } H_1 : \mu \neq 3{,}8$$

przyjmując poziom istotności $\alpha = 0{,}05$.
Znajdujemy w tablicach lub obliczamy korzystając z pakietu statystycznego (np. R-a):

$$t_{1-0{,}05/2;13} = t_{0{,}975;13} \approx 2{,}160$$

Obszar krytyczny jest więc równy:

$$(-\infty; -2{,}160] \cup [2{,}160; \infty).$$

Obliczenia— c.d.
Obliczamy t, wartość statystyki testowej T :

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3{,}581 - 3{,}8}{0.694/\sqrt{14}} \approx -1{,}18.$$

Wartość statystyki testowej t nie należy do obszaru krytycznego— nie ma podstaw do odrzucenia H_0 i przyjęcia hipotezy H_1 przy przyjętym poziomie istotności $\alpha = 0{,}05$.
p-wartość jest równa:

$$2 \times (1 - F(1{,}18)) \approx 0{,}26$$

gdzie F oznacza dystrybuantę rozkładu t-Studenta z 13 stopniami swobody.

Hipotezy alternatywne jednostronne
W przypadku, gdy hipoteza alternatywna jest prawostronna, tj. gdy testujemy

$$H_0 : \mu = \mu_0 \text{ przeciw } H_1^{'} : \mu > \mu_0$$

obszar krytyczny ma postać $[t_{1-\alpha,n-1}, \infty)$; analogicznie, gdy weryfikujemy

$$H_0 : \mu = \mu_0 \text{ przeciw } H_1^{''} : \mu < \mu_0$$

obszar krytyczny ma postać $(-\infty, -t_{1-\alpha,n-1}]$.

Przykład— obliczenia w środowisku R
Obliczamy ceny metra kw. w 14 mieszkaniach w mieście A i zapisujemy je w zmiennej `z`

```
z<-c(3.75,3.89,5.09,3.77,3.53,2.82,3.16,2.79,

4.34,3.61,4.31,3.31,2.50,3.27)
```

Przykład— obliczenia w środowisku R— c.d.
Następnie wydajemy polecenie t.test z odpowiednimi parametrami:

```
> t.test(z,mu=3.8)

        One Sample t-test

data:  z
t = -1.1777, df = 13, p-value = 0.26
alternative hypothesis: true mean is not equal to 3.8

95 percent confidence interval:
 3.180491 3.982366

sample estimates:
mean of x
 3.581429
```

Obliczenia— są wykonane dokładniej (z mniejszymi błędami zaokrągleń) niż uprzednio przedstawione rachunki;
wartość statystyki testowej $t = -1{,}1777$; df — degrees of freedom– liczba stopni swobody; p-value (p-wartośc): 0,26 95 percent confidence interval:- 95-procentowy przedział ufności- przedziałom ufności będzie poświęcony następny wykład.

Polecana literatura
Balasubramanian Narasimhan, Student's T Distribution, http://www-stat.stanford.edu/ naras/jsm/TDensity/TDensity.html
J. Koronacki i J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT 2001, str. 228–229.