MWS Ćwiczenie 4

Testy istotności

Tomasz Korzeniowski, 265753

22 grudnia 2017

1 Zadanie 1

W zadaniu rozważamy dwie hipotezy

H_0: stała intensywność samobójstw
H_1: sezonowa intensywność samobójstw

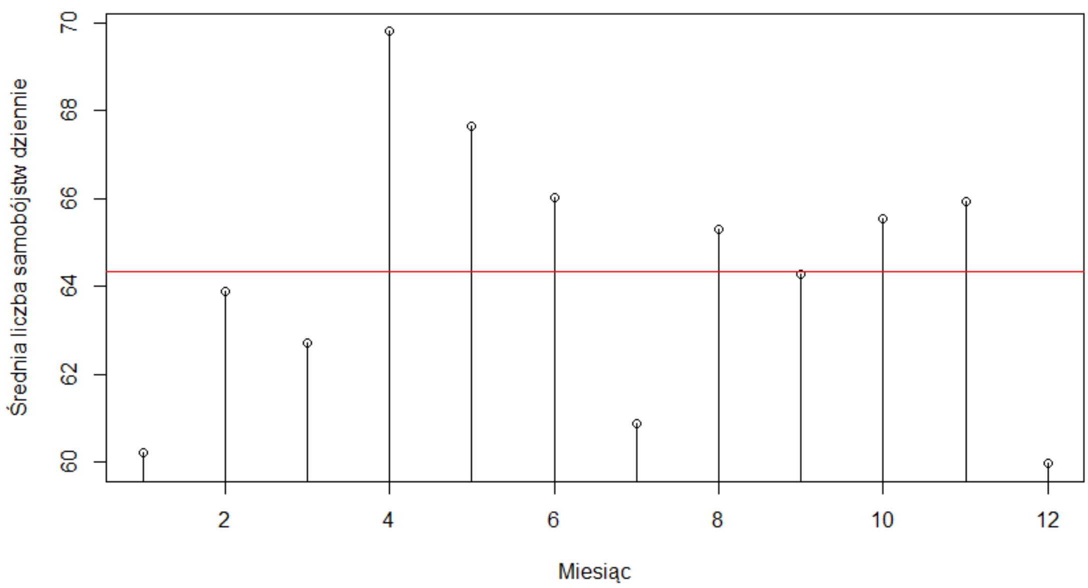
Do ich zbadania można wykorzystać twierdzenie Pearsona

$$T = \sum_{i=1}^{r} \frac{(v_i - np_i)^2}{np_i} \tag{1}$$

gdzie

r	–	liczba miesięcy
v_i	–	liczba samobójstw w i-tym miesiącu
n	–	średnia dzienna liczba samobójstw (liczona w skali całego roku)
p_i	–	liczba dni w danym miesiącu

Wiadomo, że taka statystyka zbiega do rozkładu χ^2 z (r-1) stopniami swobody. Przyjmując za poziom istotności testu $\alpha = 0,05$ otrzymujemy wartość statystyki $T = 47,36528$, wartość krytyczną $c = 19,67514$ oraz p-wartość $= 1,852011e - 06$. Na tej podstawie możemy odrzucić hipotezę zerową ($T > c$). Wynika z tego, że samobójstwa mają charakter sezonowy.



Wykres 1: Liczba samobójstw w każdym miesiącu.

Na podstawie wykresu 1 można zauważyć, że liczba samobójstw jest najmniejsza w okresie zimowym (grudzień – styczeń) oraz wakacyjnym (lipiec), a największa na wiosnę (kwiecień - maj). Zatem wynik testu potwierdza przypuszczenia o sezonowym charakterze intensywności samobójstw w ciągu roku. Czerwoną linią zaznaczono średnią liczbę samobójstw przypadającą na każdy dzień w roku.

Do wyznaczenia wartości krytycznej testu c została wykorzystana funkcja *qchisq*, której parametrami są poziom istotności $1-\alpha$ oraz liczba stopni swobody $r-1$.

Dla sprawdzenia poprawności wyników można skorzystać z funkcji *chisq.test*, która przyjmuje dane o liczbie samobójstw oraz prawdopodobieństwach ich wystąpienia (liczonych jako liczba dni w miesiącu dzielona przez liczbę dni w roku). Znalezione w ten sposób rozwiązanie daje takie same wyniki jak obliczone wcześniej:

```
        Chi-squared test for given probabilities

data:  samobojstw
X-squared = 47.365, df = 11, p-value = 1.852e-06
```

2 Zadanie 2

Estymatory największej wiarygodności rozkładu normalnego są podane wzorem

$$\hat{\mu} = \overline{X}_n, \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{2}$$

Dla temperatury ciała mężczyzn i kobiet są one następujące:

	$\hat{\mu}$	$\hat{\sigma}^2$
mężczyźni	36.72615	0.150712
kobiety	36.88923	0.170351

Tabela 1: Wartości estymatorów rozkładu normalnego temperatury ciała dla mężczyzn i kobiet.

Do przeprowadzenia testów, że średnia temperatura ciała mężczyzn lub kobiet jest równa $\mu_0 = 36.6\ ^oC$ wykorzystamy hipotezy

$$H_0\colon \mu = \mu_0$$
$$H_1\colon \mu \neq \mu_0$$

Do weryfikacji hipotez posłużymy się testem t-Studenta. Statystyka tego testu określona jest wzorem

$$TS = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \tag{3}$$

a wartość krytyczna

$$c = \Phi^{-1}(1 - \frac{\alpha}{2}) \tag{4}$$

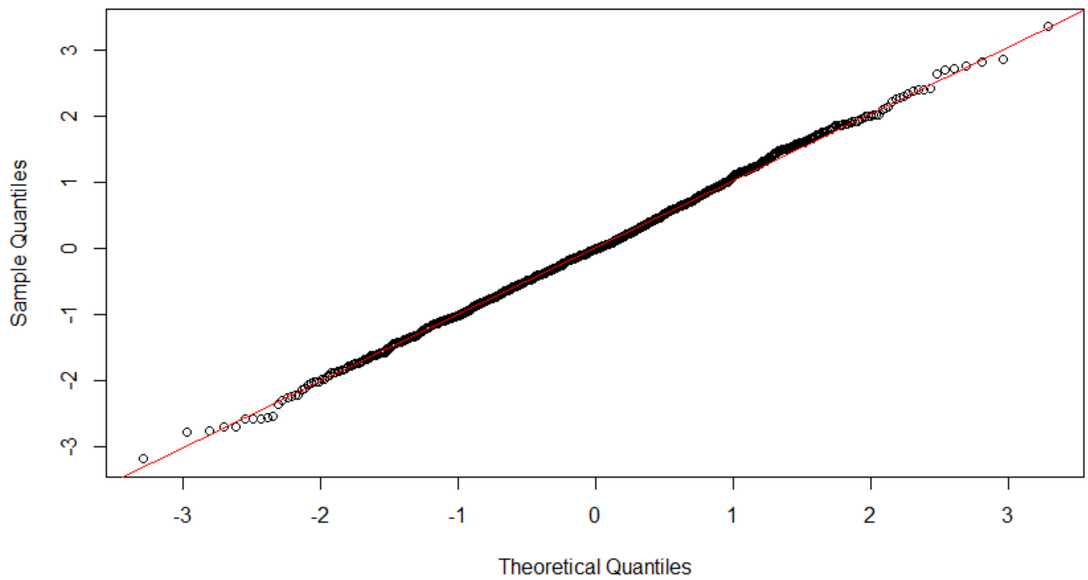
Dla danych z zadania otrzymane wyniki to

	TS	c
mężczyźni	2.619895	1.959964
kobiety	5.649745	1.959964

Tabela 2: Wartości testu studenta i wartości krytycznych dla mężczyzn i kobiet.

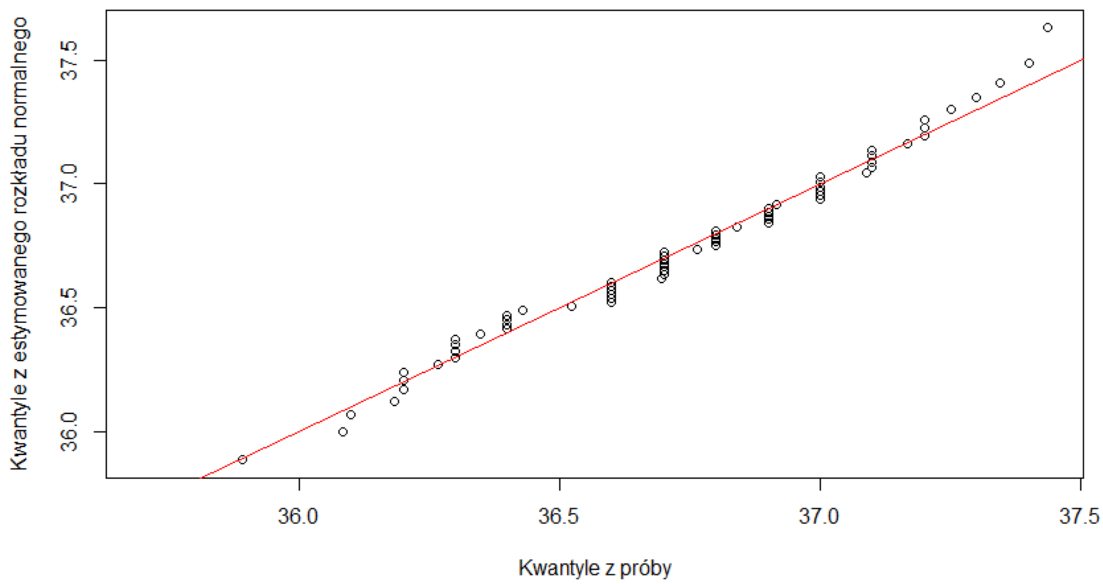
W obu przypadkach $TS > c$, zatem należy odrzucić hipotezę zerową zarówno dla mężczyzn, jak i dla kobiet.

Aby sprawdzić jak powinien wyglądać wykres kwantyl-kwantyl można wygenerować rozkład $N(0, 1)$ dla dużej liczby danych (1000). Wynik przedstawia wykres 2. Czerwoną linią zaznaczono proste, na których powinny znajdować się kwantyle należące do odpowiadających im rozkładów normalnych.



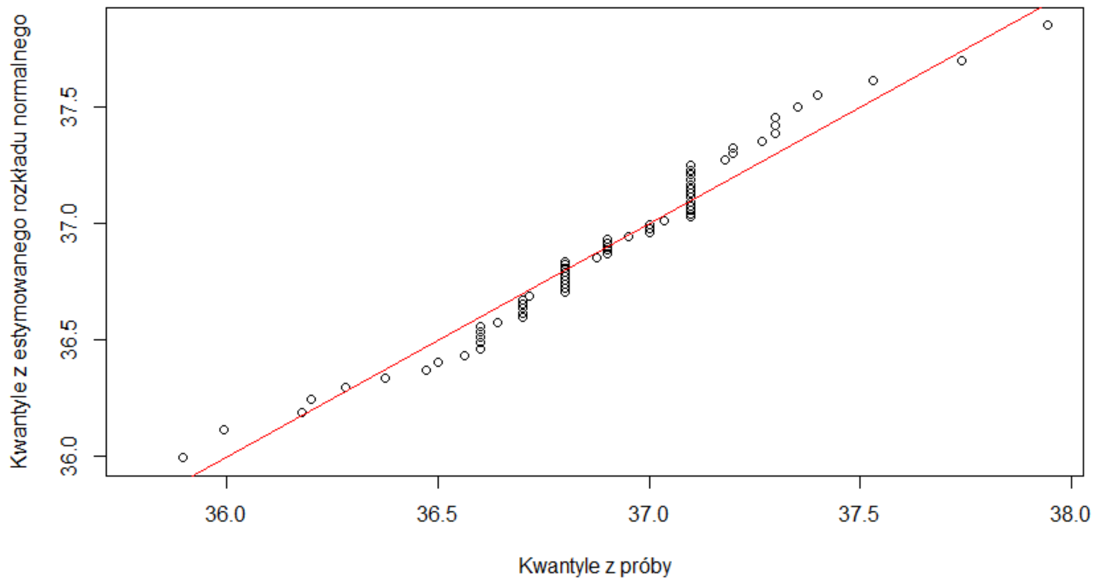
Wykres 2: Wykres kwantyl-kwantyl dla rozkładu N(0,1).

Dla danych z zadania wykresy kwantyl-kwantyl temperatur ciała obrazują wykresy 3 i 4.



Wykres 3: Wykres kwantyl-kwantyl dla temperatury mężczyzn.

Oś x wykresów przedstawia kwantyle otrzymane z próby losowej, natomiast oś y kwantyle wyliczone na podstawie rozkładu normalnego o wyestymowanych parametrach $\hat{\mu}$ i $\hat{\sigma}^2$. Można zauważyć, że dość dobrze przybliżają one rozkład normalny otrzymany jako przykład. Wraz ze



Wykres 4: Wykres kwantyl-kwantyl dla temperatury kobiet.

wzrostem liczności próby należałoby spodziewać się, że wynik będzie coraz bardziej dokładny (zbliżony do rozkładu $N(0, 1)$).

Wartości statystyk decyzyjnych zostały obliczone zgodnie z powyższymi wzorami, a do ich weryfikacji można użyć funkcji pakietu R: *t.test*. Wystarczy podać dane dla jakich ma być zweryfikowana hipoteza oraz wartość μ_0. Wyniki automatycznego testu t-Studenta dla mężczyzn

```
        One Sample t-test

data:  m$temperatura
t = 2.6199, df = 64, p-value = 0.01097
alternative hypothesis: true mean is not equal to 36.6
95 percent confidence interval:
 36.62996 36.82235
sample estimates:
mean of x
 36.72615
```

oraz dla kobiet

```
        One Sample t-test

data:  k$temperatura
t = 5.6497, df = 64, p-value = 3.985e-07
alternative hypothesis: true mean is not equal to 36.6
95 percent confidence interval:
 36.78696 36.99150
sample estimates:
mean of x
 36.88923
```