Test t cz. II

Bartosz Kozak

12 kwietnia 2019

Test t dla prób niezależnych

Test `t` dla dwóch prób niezależnych jest używany do testowania hipotezy o przynależności dwóch prób do rozkładu o tej samej średniej.

Teoria dla testu `t` dla prób niezależnych w założeniach jest bardzo podobna do teorii testu dla jednej próby. Dane pochodzą z dwóch prób: $x_{11}, x_{12}, x_{13}, ..., x_{1n}$ oraz $x_{21}, x_{22}, x_{23}, ..., x_{2n}$. Zakładamy że dane pochodzą z rozkładu $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ i hipoteza zerowa ma postać $\mu_1 = \mu_2$. Statystykę te możemy obliczyć według wzoru:

$$t = \frac{\overline{x}_2 - \overline{x}_1}{SEDM}$$

gdzie *błąd standardowy różnicy średnich* - **SEDM** obliczamy jako:

$$SEDM = \sqrt{SEM_1^2 + SEM_2^2}$$

.

Istnieją dwa algorytmy pozwalające obliczyć **SEDM**, jeden z nich zakłada równość wariancji w obu grupach (jest to tz. model "klasyczny"). Drugi natomiast dopuszcza różne wariancje w badanych grupach. W podejściu "klasycznym" obliczamy połączoną wariancję s, na podstawie odchylenia standardowego grupy 1 i 2 i podstawiamy do wzoru na **SEM**. W tym przypadku wzór na wartość t przyjmie postać:

$$t = \frac{\overline{x}_2 - \overline{x}_1}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Wariancję s^2 obliczamy stosując wzór:

$$s^2 = \frac{\sum (x_{1i} - \overline{x}_1)^2 + \sum (x_{2i} - \overline{x}_2)^2}{n_1 + n_2 - 2}$$

Wartość t będzie należeć do rozkładu `t` o $n_1 + n_2 - 2$ stopniach swobody.

Alternatywna procedura zaproponowana przez Welcha, zakłada obliczenie **SEM** z oddzielnych wariancji grup 1 i 2. W tym przypadku wartość t nie należy do dystrybucji t, ale może być przybliżone na podstawie dystrybucji `t` o liczbie stopni swobody, która może zostać obliczona na podstawie σ_1, σ_2 i wielkości grup. Zazwyczaj nie jest to liczba całkowita.

Ogólnie przyjmuje się, że procedura według Welcha jest bezpieczniejsza bo nie wymaga spełnienia warunku równości wariancji w obydwu grupach. Zazwyczaj wyniki obu wariantów testu są bardzo zbliżone (jeżeli nie ma dużych różnic między wielkością grupy i wariancją pomiędzy grupą 1 i 2).

Algorytm postępowania - test t dla grup niezależnych

1. Formułujemy hipotezę zerową (H_0) oraz alternatywną (H_A)
2. Testujemy hipotezę (wykonujemy test statystyczny)
3. Na podstawie wyników testu **odrzucamy** lub **nie odrzucamy** hipotezę zerową

Przykład 1 W pliku Cw3_1.txt znajdują się pomiary zawartość próchnicy w glebie w dwóch gospodarstwach - A oraz B. Przyjmując poziom istotności $\alpha = 0.05$ odpowiedz czy zawartość próchnicy w gospodarstwie A i B jest na takim samym poziomie.

1. Formułujemy hipotezy:

- $H_0 : \overline{x}_A = \overline{x}_B$
- $H_A : \overline{x}_A \neq \overline{x}_B$

2. Testujemy hipotezę:

- wczytujemy dane do R
- wykonujemy test t (klasyczny = zakładamy równość wariancji w grupie A i B)

3. Na podstawie wyników testu wyciągamy wnioski i udzielamy odpowiedzi na pytanie

```
# ad2
# wczytanie danych do R (plik tekstowy, dwie kolumny, wartości z nagłówkami, odzielone znakiem tabulacj
dane <- read.csv('~/konas13@gmail.com/Dokumenty/Dydaktyka/Statystyka_Ochrona_Srodowiska/Cw_3_Test_t_czI
dane
```

```
##       A   B
## 1   300 330
## 2   260 310
## 3   310 350
## 4   270 290
## 5   310 340
## 6   280 300
## 7   290 320
## 8   300 310
## 9   290 330
## 10   NA 320
```

```
# obliczamy statystykę t
# 1 obliczamy wariancję s^2
a <- dane$A[!is.na(dane$A)]
a
```

```
## [1] 300 260 310 270 310 280 290 300 290
```

```
b <- dane$B
a
```

```
## [1] 300 260 310 270 310 280 290 300 290
```

```
s2 <- (sum((a-mean(a))^2)+sum((b-mean(b))^2))/(length(a)+length(b)-2)
s2
```

```
## [1] 317.6471
```

```
# obliczamy t
t <- (mean(b)-mean(a))/sqrt((s2/length(a))+(s2/length(b)))
t
```

```
## [1] 3.663475
```

Na podstawie obliczonej statystyki t wnioskujemy na temat przyjętej hipotezy zerowej. W tym celu musimy znaleźć wartość graniczną t_α, przy $\alpha = 0.05$.

```
ta <- qt(0.95,(length(a)+length(b)-2))
ta
```

```
## [1] 1.739607
```

Na podstawie otrzymanych wyników odrzucamy H_0 na rzecz H_A, na poziomie istotności $\alpha = 0.05$, ponieważ obliczona statystyka t jest większa od wartości granicznej t_α.

```
t > ta
```

```
## [1] TRUE
```

Po przeprowadzeniu testu statystycznego możemy odpowiedzieć, że zawartość próchnicy w glebie w gospodarstwie A i B jest różna (na poziomie istotności $\alpha = 0.05$).

Możemy także wykorzystać wbudowaną funkcję R dla testu t - `t.test()`. Domyślnie funkcja t.test wykorzystuje algorytm Welcha, jeżeli chcemy użyć algorytmu "klasycznego", musimy podać parametr `var.equal = T`.

```
t.test(a,b, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  a and b
## t = -3.6635, df = 17, p-value = 0.001925
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -47.27716 -12.72284
## sample estimates:
## mean of x mean of y
##       290       320
```

W przypadku użycia funkcji t.test, wnioskujemy na podstawie *pvalue*, a nie wartości granicznej t_α. Wnioskowanie jest takie samo.

Na podstawie otrzymanych wyników odrzucamy H_0 na rzecz H_A, na poziomie istotności $\alpha = 0.05$, ponieważ *pvalue* jest mniejsza od α.

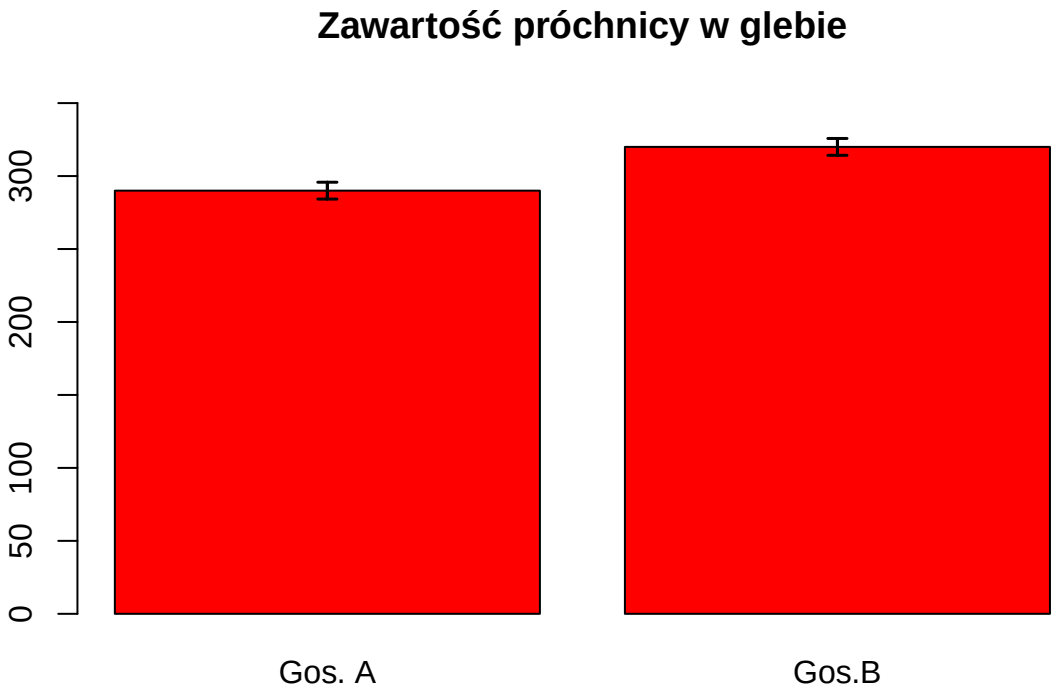
```
wynik <- t.test(a,b, var.equal = T)
wynik$p.value < 0.05
```

```
## [1] TRUE
```

Graficzna prezentacja wyników

Wyniki pomiarów zawartości próchnicy w glebie w gospodarstwie A oraz B możemy przedstawić graficznie za pomocą wykresu słupkowego `barplot`.

```
# tworzymy wektor z średnimi dla gos. A oraz B
sr <- c(mean(a),mean(b))
# tworzymy wektor z nazwami, użyjemy ich przy tworzeniu wykresu
sr_nemes <- c("Gos. A", "Gos.B")
# tworzymy wektor z wartościami  SEM, użyjemy ich przy tworzeniu słupków błędu
sr_SEM <- c(sd(a)/sqrt(length(a)),sd(b)/sqrt(length(b)))
# tworzymy wykres
wykres <- barplot(sr, names.arg = sr_nemes, ylim = c(0,350),
                  main = "Zawartość próchnicy w glebie", col = 'red')
# dodajemy do wykresu słupki błędów (SEM)
segments(wykres, sr - sr_SEM, wykres,
         sr + sr_SEM, lwd = 1.5)
arrows(wykres, sr - sr_SEM, wykres,
       sr + sr_SEM, lwd = 1.5, angle = 90,
       code = 3, length = 0.05)
```



Ćwiczenie 1

W pliku Cw3_2.txt znajduje się zestawienie plonów pszenicy z dwóch sąsiednich gmin. Na poziomie istotności $\alpha = 0.05$ odpowiedz, czy średnie plony pszenicy w tych gminach są na tym samym poziomie?

- Sformułuj H_0 oraz H_A
- Wykonaj test statystyczny
- Wyciągnij wnioski (odpowiedz na postawione pytanie)
- Przedstaw wyniki graficznie (wykres słupkowy)

Analiza danych w formacie 'stack data'

W środowisku R często spotykamy się z zapisem wyników pomiarów (danych) w "formacie stack", co oznacza, że mamy dwie kolumny. Pierwsza kolumna zawiera wartości (wyniki pomiarów), a druga kolumna informacje o przynależności pomiaru do danej grupy.

R posiada wbudowaną funkcję `stack` oraz `unstack`, która pozwala na zmianę formatu zapisu danych.

```
# zmieniamy format danych na stack
dane1 <- stack(dane)
dane1
```

```
##    values ind
## 1     300   A
## 2     260   A
## 3     310   A
## 4     270   A
## 5     310   A
## 6     280   A
## 7     290   A
## 8     300   A
## 9     290   A
## 10     NA   A
```

```
## 11    330   B
## 12    310   B
## 13    350   B
## 14    290   B
## 15    340   B
## 16    300   B
## 17    320   B
## 18    310   B
## 19    330   B
## 20    320   B
```

```r
# możemy zmienic nazwy kolumn, aby były bardziej informatywne
colnames(dane1) <- c("Zaw.pr", "Gosp")
dane1
```

```
##    Zaw.pr Gosp
## 1     300    A
## 2     260    A
## 3     310    A
## 4     270    A
## 5     310    A
## 6     280    A
## 7     290    A
## 8     300    A
## 9     290    A
## 10     NA    A
## 11    330    B
## 12    310    B
## 13    350    B
## 14    290    B
## 15    340    B
## 16    300    B
## 17    320    B
## 18    310    B
## 19    330    B
## 20    320    B
```

W przypadku danych w formacie 'stack' możemy wykonać test t korzystając z zapisu ~.

```r
t.test(Zaw.pr~Gosp, data = dane1, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  Zaw.pr by Gosp
## t = -3.6635, df = 17, p-value = 0.001925
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -47.27716 -12.72284
## sample estimates:
## mean in group A mean in group B
##             290             320
```

Test F dla wariancji

W celu sprawdzenia hipotezy o równości wariancji w badanych próbach wykorzystamy funkcje `var.test`, która implementuje test F w R. W teście tym zakładamy, losowość prób, rozkład normalny i niezależność wyników. Wzór na statystykę testową f ma postać:

$$f = \frac{s_1^2}{s_2^2}$$

Obliczoną wartość statystyki f porównujemy z wartością graniczną f_α, o liczbach stopni sfobody $n_1 - 1$ oraz $n_2 - 1$ przy zadanym *poziomie istotności*. Odrzucamy hipotezę zerową o równości wariancji w próbach, jeżeli wartość statystyki testowej f jest większa od wartości granicznej f_α. Funkcja `var.test` zwraca wartość *pvalue*, którą interpretujemy, analogicznie jak w przypadku funkcji `t.test`. Funkcję `var.test` możemy użyć na danych typu 'stack' jak i 'unstack'.

```
var.test(Zaw.pr~Gosp, data = dane1)
```

```
##
##  F test to compare two variances
##
## data:  Zaw.pr by Gosp
## F = 0.9, num df = 8, denom df = 9, p-value = 0.8926
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2194075 3.9215098
## sample estimates:
## ratio of variances
##                0.9
```

Uzyskane wyniki ($pvalue > \alpha$), nie pozwalają na odrzucenie hipotezy zerowej, więc możemy przyjąć, że zmienność zawartości próchnicy (wariancja) w gospodarstwie A i B jest taka sama.

Ćwiczenie 2

Dla danych w pliku Cw_3_1.txt obliczyć statystykę testową f i porównać ją z wartością krytyczną f_α.

Podpowiedź

Wartość krytyczną f_α możemy uzyskać z funkcji R `qf`.

Ćwiczenie 3

Czy zmienność plonów w gminie 1 i gminie 2 jest taka sama? (dane Cw3_2.txt)