Test t

Bartosz Kozak

7 kwietnia 2019

Wprowadzenie

Test `t` opiera się na założeni, że nasze dane pochodzą z rozkładu normalnego. W przypadku jednej próbki mamy dane $x_1, ..., x_n$, które powinny spełniać założenia:

- niezależność
- losowe
- pochodzą z dystrybucji $N(\mu, \sigma^2)$

Testujemy hipotezę zerową ($H_0 : \mu = \mu_0$), że nasza średnia w populacji równa jest wartości μ_0. Szacujemy wartość średnią w populacji (μ) i odchylenie standardowe populacji (σ) na podstawie średniej z próby $\overline{x}$ oraz odchylenia standardowego z próby (s).

Kluczowy koncept w tych rozważaniach to *błąd standardowy średniej*, SEM. Opisuje on zmienność (wariancję) średniej z n losowych wartości z populacji o średniej μ i wariancji σ. SEM definiujemy jako:

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

i oznacza, że jeżeli wykonamy eksperyment kilkukrotnie licząc za każdym razem średnią z otrzymanych wyników, wtedy otrzymane średnie, będą należeć do dystrybucji, która jest węższa niż oryginalna dystrybucja. Kluczową konkluzją jest to, że możemy obliczyć empiryczne SEM mając wyniki wyłącznie z jednej serii pomiarów, w takim wypadku, na podstawie empirycznej wartości s, jako $SEM = s/\sqrt{n}$. Wartość ta pozwala nam stwierdzić jak bardzo otrzymana wartość średniej (empiryczna) $\overline{x}$, może różnić się od prawdziwej wartości średniej w populacji (μ). Dla rozkładu normalnego istnieje 95% prawdopodobieństwa, że wylosowana wartość będzie w przedziale $\mu \pm 2\sigma$, możemy więc założyć, że jeżeli μ_0 jest prawdziwą średnią w populacji to $\overline{x}$ powinno znajdowac się w przedziale $2SEM$ od wartości średniej populacji (μ_0). Formalnie możemy to zapisać

$$t = \frac{\overline{x} - \mu_0}{SEM}$$

i sprawdzić czy wartość t znajduje się w przedziale prawdopodobieństwa równemu określonej przez *poziom istotności* - α.

W małych próbach, konieczne jest użycie poprawki, ponieważ w obliczeniach używamy empirycznej wartości SEM, i dystrybucja t odbiega w częściach "ogonowych" od dystrybucji rozkładu normalnego $N(0, 1)$: duże odcylenia zdażają się częściej niż w rozkładzie normalnym. Z tego powodu dla małych prób korzystamy z rozkładu $t - studenta$ o liczbie stopni swobody $f = n - 1$.

Jeżeli obliczona wartość t znajduje się poza akceptowalnym regionem przy zadymam *poziomie istotności* α wtedy odrzucamy hipotezę zerową przy tym poziomie istotności. Alternatywnie (jest to równoznaczne) możemy obliczyć wartość *p-value*, która określa prawdopodobieństwo uzyskania wartości większej od wartości t i odrzucić hipotezę (zerową) jeżeli *p-value* jest mniejsze od przyjętego *poziomu istotności*.

Przykład 1 Dane przedstawiają dzienne zapotrzebowanie energetyczne (kJ) 11 kobiet (Altman, 1991, p. 183)

```
# dane
daily.intake <- c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
```

Ćwiczenie1

Prosze obliczyć odstawowe statystyki dla danych:

- średnia

- odchylenie standardowe
- kwantyle

```
## [1] 6753.636
```

```
## [1] 1142.123
```

```
##   0%  25%  50%  75% 100%
## 5260 5910 6515 7515 8770
```

Algorytm postępowania:

1. Formułujemy hipotezę zerową H_0 i hipotezę alternatywną H_a

Należy zwrócić uwagę, że hipoteza zerowa zawsze ma postać:

$$H_0 : \mu = \mu_0$$

W naszym przypadku hipoteza zerowa będzie brzmiała: *Dzienne zapotrzebowanie energetyczne kobiet nie różni się **statystycznie** od zalecanej wartości 7725*

Hipoteza alternatywna ma postać:

$$H_a : \mu \neq \mu_0$$

2. Testujemy hipotezę (wykonujemy test statystyczny)

3. Na postawie otrzymanych wyników testu **odrzucamy** lub **nie odrzucamy** - nie mamy podstaw, hipotezę zerową na zadanym *poziomie istotności*

   Uwaga!

   Nawet jeżeli nie mamy podstaw do odrzucenia hipotezy zerowej, **nie oznacza to że jest ona prawdziwa!**

   Jeżeli odrzucimy hipotezę zerową na zadanym *poziomie istotności* oznacza to, że z prawdopodobieństwem $1 - \alpha$ hipoteza zerowa jest błędna. W takim wypadku przyjmujemy hipotezę alternatywną (z prawdopodobieństwem $1 - \alpha$).

```r
# test t w R
t.test(daily.intake,mu=7725)
```

```
##
##  One Sample t-test
##
## data:  daily.intake
## t = -2.8208, df = 10, p-value = 0.01814
## alternative hypothesis: true mean is not equal to 7725
## 95 percent confidence interval:
##  5986.348 7520.925
## sample estimates:
## mean of x
##  6753.636
```

Przedział ufności dla średniej

Jest to przedział wartości (min, max) między którymi z prawdopodobieństwem 95% znajduje się prawdziwa średnia populacji. Bazuje ona na równaniu dla statystyki t i dla poziomu istotności $\alpha = 0.05$ przyjmuje postać

$$\overline{x} - t_{0.975}(f) \times SEM < \mu < \overline{x} + t_{0.975}(f) \times SEM$$

Ćwiczenie 2

W pliku react.txt znajdują się różnice w pomiarach [obszar reakcji mm] próby tuberkulinowej na grupie pacjentów wykonane przez dwie pielęgniarki.

- Czy dane pochodzą z rozkładu normalnego?
- Na *poziomie istotności* $\alpha = 0.05$ odpowiedz na pytanie czy pomiary wykonane przez pielęgniarki różnią się istotnie statystycznie?