Testy t

Test *t* dla jednej próby

W poprzednim odcinku zajmowaliśmy się problemem pojedyńczej próby i pytaniem o populację, z której ta próba pochodzi (czy możemy odrzucić hipotezę, że średnia w populacji wynosi ileś, albo w jakim przedziale z określonym prawdopodobieństwem znajduje się średnia w populacji).

Ustaliliśmy, że test t dla jednej próby można rozumieć jako pewnego rodzaju modyfikację testu z, gdzie zamiast statystyki testowej z próby, wykorzystującej znaną wariancję w populacji:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

korzystamy ze statystyki testowej, gdzie wariancję estymujemy również z próby:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{N}}$$

Taka statystyka testowa ma przy założeniu hipotezy zerowej rozkład t o $N - 1$ stopniach swobody.

Załóżmy na przykład, że z wieloletniej praktyki wiemy, że studenci kognitywistyki rozwiązują pewin test zaliczeniowy ze statystyki ze średnim wynikiem 26.5. W tym roku pojawiło się dziesięcioro nowych studentów, którzy uzyskali `wyniki = c(10, 50, 46, 32, 37, 28, 41, 20, 32, 43)`. Czy mamy podstawy sądzić, że ci studenci są lepsi ze statystyki niż przeciętny student kognitywistyki? Odpowiedzieć na to pytanie pomoże nam test t dla jednej próby: `t.test(wyniki, mu=26.5)`.

Pytanie

Które z poniższych twierdzeń są prawdziwe?

- Prawdopodobieństwo wylosowania z populacji o średniej 26.5 dziesięcioelementowej próby o średniej co najmniej tak różnej od 26.5 jak obserwowana średnia 33.9 wynosi około 8.9%.
- Zakładając $\alpha = 0.05$, nie mamy podstaw, żeby odrzucić hipotezę zerową, że tych dziesięcioro studentów pochodzi z populacji o średniej 26.5.
- Ponieważ 95-procentowy przedział ufności wokół średniej w próbie obejmuje wartość 26.5, nie mamy podstaw, żeby odrzucić hipotezę, że tych dziesięcioro studentów pochodzi z populacji o średniej 26.5.
- Średni wynik w próbie wynosi 33.9.

Odpowiedź

Wszystkie cztery są prawdziwe.

Test *t* dla prób zależnych

W praktyce badawczej znacznie częściej mamy do czynienia z dwiema próbami, a nie jedną. W takiej sytuacji hipoteza zerowa mówi zazwyczaj, że próby te pochodzą z populacji o tej samej średniej. Czy kobiety mają wyższą inteligencję werbalną od mężczyzn? Czy wypicie filiżanki kawy zmniejsza czas reakcji w jakimś zadaniu poznawczym? Czy przestawienie szyku zdania powoduje, że wolniej to zdanie rozumiemy (Człowiek pogryzł psa a Psa pogryzł człowiek)? To są wszystko przykłady problemów z taką właśnie hipotezą zerową.

Szczególnym przypadkiem jest sytuacja, kiedy dwie próby są ze sobą powiązane w taki sposób, że pomiary z obu prób łączą się w pary. Na przykład losowej grupie studentów kognitywistyki mierzymy inteligencję przed początkiem semestru i ponownie na początku ferii zimowych, bo naszym pytaniem badawczym jest, czy 30 godzin statystyki z R w ciągu semestru podnosi inteligencję ogólną. Mamy zatem teoretycznie dwie próby pomiarów, ale tylko o tej pierwszej moglibyśmy powiedzieć, że była losowa: druga jest w całości zdeterminowana tą pierwszą (bo mierzymy inteligencję ponownie tym samym osobom).

W takiej sytuacji stosujemy test t dla prób zależnych (po angielsku brzmi to nawet lepiej, bo *paired-sample t test*), który w istocie nie różni się niczym od testu t dla jednej próby. Interesują nas pary pomiarów, a właściwie różnice między dwoma pomiarami w każdej parze. Hipoteza zerowa mówi, że średnia taka różnica w populacji wynosi zero.

No to załóżmy, że przetestowaliśmy w ten sposób 49 studentów kognitywistyki, każdemu mierząc inteligencję przed i po kursie statystyki z R i odejmując pierwszy pomiar od drugiego. Niektórzy wypili więcej kawy przed pierwszym pomiarem, inni byli zmęczeni sesją egzaminacyjną, więc niektóre różnice wyszły ujemne, a średnia różnica wyniosła tylko pół punkta (na korzyść późniejszego pomiaru).

Pytanie I

Jaka jest wartość statystyki t przy założeniu, że odchylenie standardowe tych 49 różnic wyniosło aż 1.75?

Wartość t otrzymujemy, dzieląc różnicę między średnią w próbie (0.5) a zakładaną w hipotezie zerowej średnią w populacji (0) przez odchylenie standardowe w próbie (1.75) dzielone przez pierwiastek liczebności (49). Następnie możemy skorzystać z rozkładu t o $N - 1$ stopniach swobody, by oszacować prawdopodobieństwo uzyskania co najmniej tak skrajnej różnicy między średnią w próbie a średnią w populacji, jak ta, którą uzyskaliśmy.

Pytanie II

Jaka będzie nasza konkluzja?

- Przyjmując $\alpha = 0.01$, nie mamy podstaw, żeby odrzucić hipotezę zerową, że 30 godzin statystyki w semestrze nie wpływa na poziom inteligencji ogólnej.
- Nawet gdybyśmy całkowicie odrzucili możliwość, że wpływ kursu statystyki na inteligencję może być negatywny, i przeprowadzili test jednostronny (czyli zakładający kierunkową hipotezę alternatywną), przy $\alpha = 0.01$ nie mielibyśmy podstaw, żeby odrzucić hipotezę zerową.
- Przyjmując $\alpha = 0.05$, nie mamy podstaw, żeby odrzucić hipotezę zerową, że 30 godzin statystyki w semestrze nie wpływa na poziom inteligencji ogólnej.
- Nawet gdybyśmy całkowicie odrzucili możliwość, że wpływ kursu statystyki na inteligencję może być negatywny, i przeprowadzili test jednostronny (czyli zakładający kierunkową hipotezę alternatywną), przy $\alpha = 0.05$ nie mielibyśmy podstaw, żeby odrzucić hipotezę zerową.

Odpowiedzi

- Wartość statystyki testowej t wynosi 2!
- Fałszywa jest opcja numer 4

Test *t* dla prób zależnych

Jeszcze jeden przykład. Amerykańscy badacze porównywali satysfakcję z życia seksualnego kobiet i mężczyzn żyjących w stałym związku (badanie jest autentyczne, ale liczby zmyśliłem). Załóżmy, że w tym celu przepytali 15 par i każda osoba musiała ocenić swoją satysfakcję na skali 1–5. Mamy więc 15 odpowiedzi mężczyzn i 15 odpowiedzi kobiet, ale zasadne jest analizowanie ich parami.

Dlaczego? Przyjmijmy, że na poziomie populacji średnia satysfakcja mężczyzn jest wyższa od średniej kobiet. Mimo to w naszej próbie mógł się trafić mężczyzna, który jest na tyle nieusatysfakcjonowany, że będzie wyjątkowo zaniżał średnią w grupie mężczyzn (a zatem "utrudniał" odkrycie istniejącej na poziomie populacji różnicy). Może to być jednak spowodowane obiektywnym faktem, że życie seksualne danej pary nie układa się najlepiej, a ocena tego mężczyzny, mimo że niska na tle innych mężczyzn, może wciąż przewyższać ocenę jego partnerki. Stosując test t dla prób zależnych, bierzemy właśnie to pod uwagę.

Jeśli odpowiedzi mężczyzn są takie:

```
m <- c(1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5)
```

a odpowiedzi kobiet takie:

```
k <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5)
```

zakładając, że odpowiadające sobie elementy obu wektorów tworzą pary, możemy zastosować `t.test(m, k, paired=TRUE)`.

Pytanie

Wybierz prawidłową odpowiedź:

- 95-procentowy przedział ufności nie obejmuje 0. Wynika z tego, że przy $\alpha = 0.05$ możemy odrzucić hipotezę zerową.
- 95-procentowy przedział ufności nie obejmuje 0. Wynika z tego, że przy $\alpha = 0.01$ możemy odrzucić hipotezę zerową.
- 95-procentowy przedział ufności nie obejmuje 0. Wynika z tego, że przy $\alpha = 0.05$ nie możemy odrzucić hipotezy zerowej.
- 95-procentowy przedział ufności nie obejmuje 0. Wynika z tego, że przy $\alpha = 0.01$ nie możemy odrzucić hipotezy zerowej.

Odpowiedź

Oczywiście prawidłową odpowiedzią jest odpowiedź pierwsza.

Test *t* dla prób zależnych - przypomnienie.

Amerykańscy badacze porównywali satysfakcję z życia seksualnego kobiet i mężczyzn żyjących w stałym związku (badanie jest autentyczne, ale liczby zmyśliłem). Załóżmy, że w tym celu przepytali 15 par i każda osoba musiała ocenić swoją satysfakcję na skali 1–5. Mamy więc 15 odpowiedzi mężczyzn i 15 odpowiedzi kobiet, ale zasadne jest analizowanie ich parami.

Dlaczego? Przyjmijmy, że na poziomie populacji średnia satysfakcja mężczyzn jest wyższa od średniej kobiet. Mimo to w naszej próbie mógł się trafić mężczyzna, który jest na tyle nieusatysfakcjonowany, że będzie wyjątkowo zaniżał średnią w grupie mężczyzn (a zatem "utrudniał" odkrycie istniejącej na poziomie populacji różnicy). Może to być jednak spowodowane obiektywnym faktem, że życie seksualne danej pary nie układa się najlepiej, a ocena tego mężczyzny, mimo że niska na tle innych mężczyzn, może wciąż przewyższać ocenę jego partnerki. Stosując test *t* dla prób zależnych, bierzemy właśnie to pod uwagę.

Jeśli odpowiedzi mężczyzn są takie:

```
m <- c(1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5)
```

a odpowiedzi kobiet takie:

```
k <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5)
```

zakładając, że odpowiadające sobie elementy obu wektorów tworzą pary, możemy zastosować `t.test(m, k, paired=TRUE)`.

```
m <- c(1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5)
k <- c(1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5)

t.test(m,k, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  m and k
```

```
## t = 4, df = 14, p-value = 0.001316
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.2473618 0.8193049
## sample estimates:
## mean difference
##       0.5333333
```

Test *t* dla prób niezależnych

Różnica między kobietami a mężczyznami okazała się istotna statystycznie. Wartość *p* była niższa niż .05, a 95-procentowy przedział ufności nie objął zera. Co by jednak było, gdybyśmy te same odpowiedzi uzyskali od 15 kobiet i 15 mężczyzn wylosowanych zupełnie niezależnie od siebie?

Nie moglibyśmy odpowiedzi połączyć w pary i zastosować tej samej procedury, co w przypadku pojedyńczej próby. W przypadku pojedyńczej próby punktem wyjścia jest rozkład średniej z próby, czyli rozkład określający prawdopodobieństwo wylosowania prób o różnych średnich z populacji o danej średniej (założonej w hipotezie zerowej). Podobnie jest w przypadku dwóch prób zależnych, kiedy interesują nas różnice w parach (a hipotetyczna średnia tych różnic w populacji wynosi zero). Natomiast w przypadku dwóch prób niezależnych statystyką, która nas interesuje, nie jest pojedyńcza średnia, tylko różnica między średnimi.

Przede wszystkim więc musimy określić rozkład różnicy średnich z prób. Jak wygląda taki rozkład? Na podstawie naszej wiedzy statystycznej możemy już formułować pewne intuicje. Jako coraz bardziej zaawansowani użytkownicy R możemy z kolei nasze intuicje sprawdzać!

```
N <- 20
mu <- 165
sigma1 <- 10
sigma2 <- 8
i <- 100000
s1 <- replicate(i, mean(rnorm(N, mu, sigma1)))
s2 <- replicate(i, mean(rnorm(N, mu, sigma2)))
```

Pytanie

`s1` i `s2` to przybliżenia rozkładu średniej z próby, każde bazujące na `i` losowań z populacji o tej samej średniej mu. Zatem `s1 - s2` będzie przybliżeniem rozkładu różnicy średnich z prób.

Które z poniższych stwierdzeń są prawdziwe?

- `var(s1)` oscyluje wokół `sigma1^2/N`.
- `sd(s2)` oscyluje wokół `sigma2/sqrt(N)`.
- `var(s1-s2)` oscyluje wokół `var(s1) + var(s2)`.
- `mean(s1-s2)` oscyluje wokół `mu * 2`.

```
var(s1)
```

```
## [1] 4.977071
```

```
sd(s2)
```

```
## [1] 1.782954
```

```
var(s1-s2)
```

```
## [1] 8.15572
```

```
mean(s1-s2)
```

```
## [1] 0.009977828
```

Rozkład różnicy średnich z prób to rozkład normalny o średniej równej różnicy średnich w populacji (zazwyczaj, zgodnie z hipotezą zerową, 0).

Wariancja tego rozkładu jest natomiast sumą wariancji rozkładów poszczególnych średnich.

Wariancja rozkładu średniej z próby to w pewnym sensie niedokładność, z jaką średnia próby odpowiada średniej w populacji, a związana z losowaniem tej próby. Jeśli losujemy z tej samej populacji dwie próby, to różnica pomiędzy średnimi powinna oczywiście oscylować wokół zera, ale tak jak nie możemy być pewni, że średnia jednej czy drugiej próby jest równa średniej w populacji, tym bardziej nie możemy być pewni, że różnica tych średnich będzie równa zero.

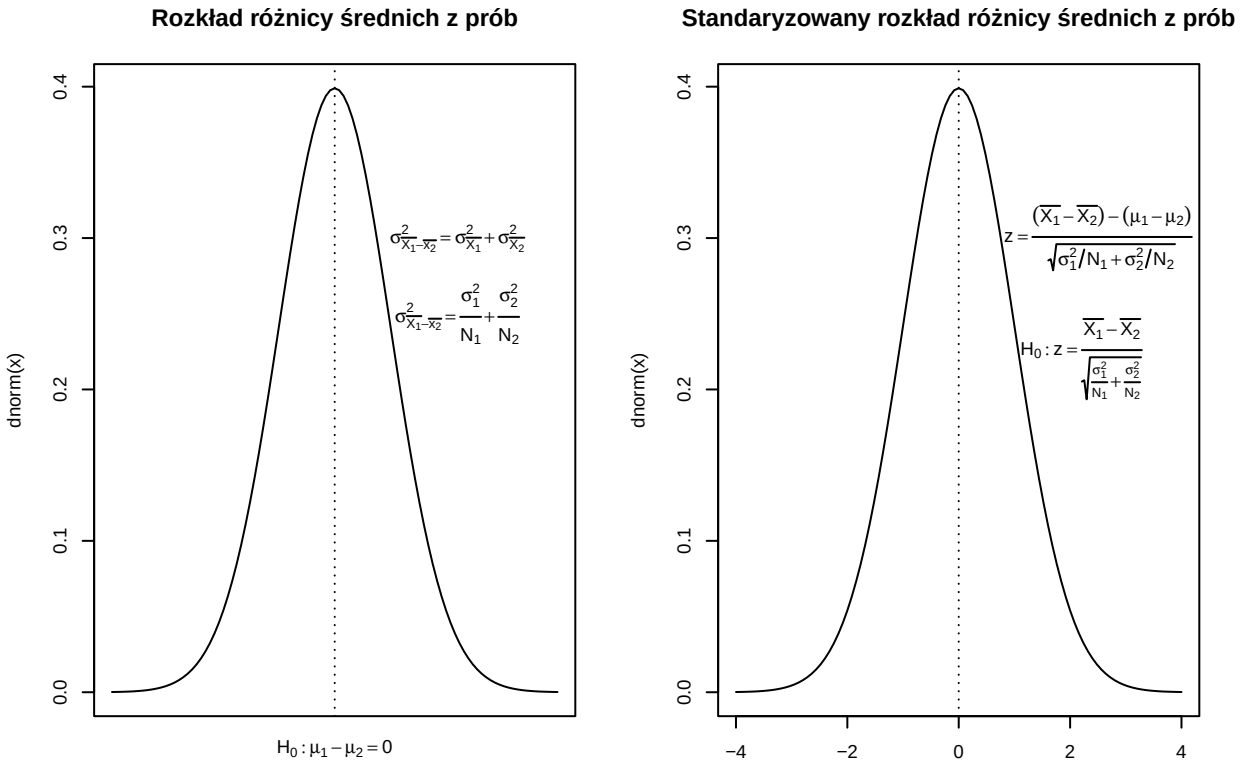
```
options(repr.plot.width=7.5, repr.plot.height=5)

par(mfrow = c(1,2), cex = 0.6)

wariancja_roznicy <- expression(sigma[bar(X[1]) - bar(x[2])]^2 == sigma[bar(X[1])]^2 + sigma[bar(X[2])]
wariancja_roznicy2 <- expression(sigma[bar(X[1]) - bar(x[2])]^2 == frac(sigma[1]^2, N[1]) + frac(sigma[
curve(dnorm(x),
     main = 'Rozkład różnicy średnich z prób',
     from= -4,
     to=4,
     xaxt = 'n',
     xlab = '')
text(2.2, 0.3, labels = wariancja_roznicy)
text(2.2, 0.25, labels = wariancja_roznicy2)
abline(v = 0, lty = 3)
mtext(side = 1, expression(H[0]: mu[1] - mu[2] == 0), cex = 0.6, line = 1)

z_roznicy <- expression(z == frac((bar(X[1]) - bar(X[2])) - (mu[1] - mu[2]),
                                  sqrt(sigma[1]^2/N[1] + sigma[2]^2/N[2])))
h0_z <- expression(H[0]: z == frac(bar(X[1]) - bar(X[2]),
                      sqrt(frac(sigma[1]^2, N[1]) + frac(sigma[2]^2, N[2]))))

curve(dnorm(x),
     main = 'Standaryzowany rozkład różnicy średnich z prób',
     from= -4,
     to=4,
     xlab = '')
text(2.5, 0.3, labels = z_roznicy)
text(2.2, 0.22, labels = h0_z)
abline(v = 0, lty = 3)
```



Pozostaje jeszcze pytanie, czy zamieniając we wzorze na z (drugi wykres) obie sigmy (σ, których zazwyczaj nie znamy) na wariancje policzone w próbach (s), otrzymamy statystykę o rozkładzie t, jak w przypadku pojedyńczej próby. William Gosset (a.k.a. Student), który w ogóle jako pierwszy zaproponował rozkład t, wykazał, że tak, pod warunkiem, że obie wartości s są estymatorami tego samego parametru (czyli, że wariancje w obu populacjach są identyczne: nie jest to założenie pozbawione sensu, a poza tym pojawia się we wszystkich pokrewnych metodach, jak analiza wariancji, czy analiza regresji; warto więc o nim pamiętać). Gosset zaproponował, żeby w takim wypadku zamiast dwóch oddzielnych estymatorów stosować jeden uśredniony (licząc ich średnią ważoną, czyli biorąc poprawkę na różnicę w liczebnościach obu prób). Rozkład prawdopodobieństwa otrzymanej statystyki to rozkład t o liczbie stopni swobody równej sumie stopni swobody w obu próbach.

Pytanie

Zatem liczba stopni swobody w teście t Studenta dla prób niezależnych wynosi:

- $N_1 + N_2 - 2$
- $N_1 + N_2 - 1$
- $N - 1$
- $2N$

Problem Behrensa-Fishera

Pytanie o to, jaki rozkład ma statystyka t w przypadku heterogeniczności wariancji (czyli, gdy wariancje w obu populacjach są różne), nazywa się problemem Behrensa-Fishera. Najczęściej używanym przybliżonym rozwiązaniem tego problemu jest rozwiązanie Welcha-Satterthwaite'a, które zakłada, że jest to rozkład t, ale o liczbie stopni swobody mniejszej niż $N_1 + N_2 - 2$. Dokładną liczbę *df* daje rozwiązanie równania Welcha-Satterthwaite'a, a test t stosowany z tą poprawką nazywa się często testem t Welcha.

Pytanie

Jeśli wynik testu t Studenta jest istotny statystycznie, czy test t Welcha policzony na tych samych danych też okaże się istotny (zakładając ten sam poziom α)?

- Niekoniecznie. Jest raczej na odwrót: jeśli test t Welcha jest istotny statystycznie, test t Studenta policzony na tych samych danych również będzie istotny.
- Tak.
- Niekoniecznie. Zależy od tego, czy mamy do czynienia z heterogenicznością wariancji.
- Niekoniecznie.

Test *t* Studenta vs test *t* Welcha

Kiedy stosować test t Studenta, a kiedy — test t Welcha? Najlepiej zawsze Welcha (i tak domyślnie robi R), bo jest to rozwiązanie konserwatywne: jeśli uzyskamy istotny statystycznie wynik w teście Welcha, to pewnie tym bardziej uzyskalibyśmy taki wynik w teście Studenta. Podstawowa różnica to redukcja *df* względem $N_1 + N_2 - 2$ w rozwiązaniu Welcha, a im mniejsza liczba df, tym grubsze ogony rozkładu t i tym dalej zaczyna się obszar krytyczny: tym trudniej w niego "wpaść".

Pytanie

Na koniec wróćmy do naszych mężów i żon. Jaki wynik uzyskamy, jeśli przyjmiemy, że nie byli to mężowie i żony, ale 15 kobiet i 15 mężczyzn wylosowanych do badania niezależnie od siebie?

- Wartość t w teście Welcha wynosi 1.0583.
- Liczba stopni swobody w teście Studenta wynosi 28.
- Wartość t w teście Studenta wynosi 1.0237.
- Liczba stopni swobody w teście Welcha wynosi 28.

Te same dane prowadzą do zupełnie innych wniosków, jeśli potraktować je innym (niewłaściwym) testem. Wartość statystyki t w teście dla prób niezależnych zawsze będzie niższa, bo mianownik jest większy: skoro obie próby losowane były niezależnie, to musimy uwzględnić (dodać do siebie) wariancje generowane przez nie obie.