Laboratorium nr 8

Statystyka matematyczna rok ak. 2023/24

ZADANIE 8.1. Wysunięto przypuszczenie, że palenie tytoniu może powodować powstawanie zmarszczek na skórze wokół oczu. Zbadano, więc 150 palaczy i 250 osób niepalących i stwierdzono, że u 95 palaczy i 105 osób niepalących zaobserwowano widoczne zmarszczki wokół oczu (na podstawie standardowej oceny zmarszczek przeprowadzonej przez osobę, która nie wiedziała, czy badana osoba jest paląca, czy nie).

(A) Napisz funkcję programu R za pomocą której można zbudować przedział ufności dla frakcji.

Przykładowa funkcja wyznaczająca przedział ufności dla frakcji.

```
pu.frak <- function(k=k,n=n,alfa=alfa){
a <- 1-alfa
u.a <- qnorm(1-a/2)
frak <- k/n
l.kon <- round(frak-u.a*sqrt(frak/n*(1-frak)),3)
p.kon <- round(frak+u.a*sqrt(frak/n*(1-frak)),3)
przedzial <- paste('(',l.kon,',',p.kon,')')
return(przedzial)
}
```

(B) Korzystając ze budowanej funkcji w punkcie (A) zbuduj przedziały ufności dla frakcji osób posiadających zmarszczki wokół oczu w przypadku osób palących i niepalących. Przyjmij poziom ufności 0.95. Czy na podstawie zbudowanych przedziałów możesz sformułować jakieś wnioski?

Przedział ufności dla frakcji osób ze zmarszczkami wokół oczu wśród palących

```
pu.frak(95,150,0.95)
```

```
## [1] "( 0.556 , 0.71 )"
```

Przedział ufności dla frakcji osób ze zmarszczkami wokół oczu wśród niepalących

```
pu.frak(105,250,0.95)
```

```
## [1] "( 0.359 , 0.481 )"
```

Zauważmy, że otrzymane przedziały są rozłączne. Wobec tego, możemy przypuszczać, że na podstawie otrzymanych prób oraz na poziomie istotności 0.95, istnieje różnica między osobami palącymi tytoń i niepalącymi w kontekście posiadania zmarszczek wokół oczu.

(B) Aby stwierdzić czy palenie tytoniu może powodować powstawanie zmarszczek na skórze wokół oczu zbuduj 95% przedział ufności dla różnicy frakcji osób ze zmarszczkami wokół oczu wśród palących i niepalących. Jaki wniosek możesz sformułować na podstawie otrzymanego przedziału?

Skorzystamy ze wzoru

$$\left(\hat{p}_1 - \hat{p}_2 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\right.$$

$$\hat{p}_1 - \hat{p}_2 + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\Bigg)$$

gdzie $\hat{p}_1$ jest estymatorem punktowym frakcji osób palących posiadających zmarszczki wokół oczu, a $\hat{p}_2$ – osób niepalących. Użyjemy poniższego kodu

```
p_1 <- 95/150
p_2 <- 105/250
u.a <- qnorm(1-0.05/2)
l.kon <- p_1-p_2-u.a*sqrt(p_1*(1-p_1)/150+p_2*(1-p_2)/250)
p.kon <- p_1-p_2+u.a*sqrt(p_1*(1-p_1)/150+p_2*(1-p_2)/250)
```

otrzymując następujący przedział ufności (0.115, 0.312).

Otrzymany przedział ufności nie pokrywa zera, co oznacza, że istnieje istotna różnica między osobami palącymi i niepalącymi, na poziomie ufności 0.95 i na podstawie otrzymanych prób. Ponadto widzimy, że oba końce przedziału są dodatnie, co może świadczyć o tym, że $p_1 > p_2$, czyli frakcja osób palących posiadających zmarszczki wokół oczu jest większa.

ZADANIE 8.2. Na laboratorium nr 3 rozważaliśmy zbiór danych *stenzenieolowiu*. W zbiorze znajdują się dwie zmienne: stężenie ołowiu w wodzie znane i zmierzone. Na podstawie tych danych chcemy stwierdzić, czy metoda, którą dokonano pomiarów jest dobrze wyskalowana.

Na poziomie ufności 0.95 zbuduj przedział ufności dla różnicy średnich stężeń znanego i zmierzonego, a następnie na podstawie otrzymanego przedziału odpowiedz na pytanie, czy rozważana metoda jest dobrze wyskalowana. Obliczeń dokonaj korzystając z odpowiednich wzorów podanych na wykładzie, a następnie za pomocą funkcji *stats::t.test()*.

Ponieważ badamy stężenie ołowiu zmierzone i znane, więc dane potraktujemy jako pary obserwacji pochodzące z tej samej populacji dwuwymiarowej. Zbudujemy przedział ufności dla różnicy stężeń. W tym celu utworzymy nową próbę różnic między stężeniem znanym a zmierzonym

```
roz.st <- stezenie$znane-stezenie$zmierzone

n <- length(roz.st)

alfa <- 1-0.95
x.bar <- mean(roz.st)
s.hat <- sd(roz.st)
kw <- qt(1-alfa/2,n-1)
l.kon <- round(x.bar-kw*s.hat/sqrt(n),3)
p.kon <- round(x.bar+kw*s.hat/sqrt(n),3)
```

Przedział ufności (-0.039, 0.594), na poziomie ufności 0.95 pokrywa nieznaną różnicę stężeń. Ponieważ otrzymany przedział pokrywa zero, więc możemy twierdzić, że nie ma istotnej różnicy miedzy stężeniem znanym a zmierzonym, co oznacza, że metoda pomiaru stężenia ołowiu jest dobrze wyskalowana.

Budowa przedziału ufności za pomocą funkcji *stats::t.test()*

```
t.test(roz.st)
```

```
##
##  One Sample t-test
##
## data:  roz.st
## t = 1.9277, df = 11, p-value = 0.0801
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.03934186  0.59434186
```

```
## sample estimates:
## mean of x
##    0.2775
```

Obliczeń można dokonać również za pomocą polecenia

```
t.test(stezenie$znane,stezenie$zmierzone,paired = T)
```

```
##
##  Paired t-test
##
## data:  stezenie$znane and stezenie$zmierzone
## t = 1.9277, df = 11, p-value = 0.0801
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.03934186  0.59434186
## sample estimates:
## mean difference
##          0.2775
```

ZADANIE 8.3. W celu porównania zawartości kalorii w hot-dogach wołowych i drobiowych zmierzono zawartość kalorii w 20 hot-dogach wołowych i 17 drobiowych. Wyniki zapisano w zbiorze *hot_dog* udostępnionym na Teamsach. Zakładając, że dane pochodzą z populacji o rozkładach normalnych, na poziomie ufności 0.98 zbuduj przedział ufności dla różnicy średniej zawartości kalorii w rozważanych hod-dogach. Obliczeń dokonaj korzystając z odpowiednich wzorów podanych na wykładzie w dwóch przypadkach

- zakładając, że odchylenia standardowe rozkładów kalorii w rozważanych hod-dogach są równe,
- zakładając, że odchylenie standardowe rozkładów kalorii rozważanych hod-dogach są różne.

Czy na podstawie zbudowanych przedziałów możesz sformułować jakieś wnioski?

Na początku wczytamy dane

```
library(readxl)
hotdog <- read_xlsx('D:\\SM_IAD_2324\\dane\\hot_dog.xlsx')
h.w <- hotdog$wołowe
h.d <- hotdog$drobiowe[1:17]
```

Sprawdzamy założenia

- dane pochodzą z populacji o rozkładach normalnych,
- liczebności prób są mniejsze od 30,

Następnie załóżmy, ze nieznane odchylenia standardowe rozkładu kalorii rozważanych hod-dogach są **równe**. Skorzystamy, więc ze wzoru

$$\left((\overline{x}-\overline{y}) - t(1-\frac{\alpha}{2}, n_1+n_2-2)s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\right.$$

$$\left.(\overline{x}-\overline{y}) + t(1-\frac{\alpha}{2}, n_1+n_2-2)s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right),$$

gdzie

$$s_p^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}$$

Posłużymy się następującym kodem

```
n_1 <- 17
n_2 <- 20
x.bar <- mean(h.d)
y.bar <- mean(h.w)
s_1 <- sqrt(sum((h.d-x.bar)^2)/(n_1-1))
s_2 <- sqrt(sum((h.w-y.bar)^2)/(n_2-1))
s.p <- sqrt(s_1^2*(n-1)+s_2^2*(n-1))/sqrt(n_1+n_2-2)
a <- 1-0.98
t.a <- qt(1-a/2,n_1+n_2-2)
l.kon <- round(x.bar-y.bar-t.a*s.p*sqrt(1/n_1+1/n_2),3)
p.kon <- round(x.bar-y.bar+t.a*s.p*sqrt(1/n_1+1/n_2),3)
```

otrzymując przedział ufności (-49.748, -19.011).

Zauważmy, że końce obu przedziałów są ujemne. Ponieważ w powyższym kodzie użyliśmy różnicy między zawartością kalorii w hot-dogach drobiowych a wołowych, to na poziomie ufności 0.98 możemy przypuszczać, że ta różnica jest ujemna, czyli liczba kalorii w hod-dogach drobiowych jest mniejsza niż w wołowych.

Załóżmy teraz, że odchylenia standardowe rozkładu kalorii rozważanych hod-dogach są **różne**. Skorzystamy, więc ze wzoru

$$\left((\overline{x}-\overline{y}) - t(1-\frac{\alpha}{2},\nu)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\ (\overline{x}-\overline{y}) + t(1-\frac{\alpha}{2},\nu)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \right)$$

gdzie

$$\nu = \left\lceil \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1-1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2-1}} \right\rceil .$$

Skorzystamy więc z kodu

```
s_1 <- sqrt(sum((h.d-x.bar)^2)/(n_1-1))
s_2 <- sqrt(sum((h.w-y.bar)^2)/(n_2-1))
nu <- ceiling((s_1^2/n_1+s_2^2/n_2)^2/((s_1^2/n_1)^2/(n_1-1)
              +(s_2^2/n_2)^2/(n_2-1)))
t.a <- qt(1-a/2,nu)

l.kon <- round(x.bar-y.bar-t.a*sqrt(s_1^2/n_1+s_2^2/n_2),3)
p.kon <- round(x.bar-y.bar+t.a*sqrt(s_1^2/n_1+s_2^2/n_2),3)
```

otrzymując przedział ufności (-53.912, -14.847).