ADPS 2020Z — Laboratorium 3 (rozwiązania)

Jakub Adamowicz

Zadanie 1

Treść zadania

Plik tempciala.txt zawiera zarejestrowane wartości tętna oraz temperatury ciała dla 65 mężczyzn (płeć = 1) i 65 kobiet (płeć = 2).

Osobno dla mężczyzn i kobiet:

- wyestymuj wartość średnią i odchylenie standardowe temperatury,
- zweryfikuj hipotezę, że średnia temperatura jest równa 36.6 °C wobec hipotezy alternatywnej, że średnia temperatura jest inna, przyjmując, że temperatury mają rozkład normalny,
- przeprowadź testy normalności dla zarejestrowanych temperatur.

Rozwiązanie

Estymacja wartości średniej i odchylenia standardowego

```
temp_ciala = read.csv('tempciala.txt')

man_temp = temp_ciala[temp_ciala$płeć == 1,]
woman_temp = temp_ciala[temp_ciala$płeć == 2,]

mean_est_man = mean(man_temp$temperatura)
sd_est_man = sd(man_temp$temperatura)

mean_est_woman = mean(woman_temp$temperatura)
sd_est_woman = sd(woman_temp$temperatura)
```

Wyestymowana wartość średnia temperatury dla kobiet wynosi 36.89, a wyestymowane odchylenia standardowe wynosi 0.41.

Wyestymowana wartość średnia temperatury dla mężczyzn wynosi 36.73, a wyestymowane odchylenia standardowe wynosi 0.39.

Założenie hipotezy zerowej H_0 i alternatywnej H_1 - dotyczy zarówno danych dotyczących kobiet jak i mężczyzn.

$$H_0 : \mu = 36.6,$$
$$H_1 : \mu \neq 36.6,$$

```
mi_0 = 36.6
alfa = 0.05
```

Poziom istosności α wynosi 0.05.

Obliczenia dla temparatury ciała kobiet

Weryfikacja hipotezy z założeniem, że wariancja nie jest znana

```r
n_woman = length(woman_temp$temperatura)

T_woman = abs(mean_est_woman -mi_0)*sqrt(n_woman)/sd_est_woman
c_woman = qt(1-alfa/2,df = n_woman-1)
p_val_woman = 2*(1 -pt(T_woman, df = n_woman-1))
```

Wartość statystyki T = 5.6497454. Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 1.9977297.

p-wartość = 3.9852716×10^{-7}.

Dla zadanej wartości α hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną.

Wykorzystanie funkcji t.test:

```r
t.test(woman_temp$temperatura, mu = mi_0, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  woman_temp$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
```

Dla zadanej wartości α hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną.

Za pomocą testu Shapiro-Wilka zweryfikuj hipotezę, że dane pochodzą z rozkładu normalnego:

```r
shapiro.test(woman_temp$temperatura)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  woman_temp$temperatura
## W = 0.95981, p-value = 0.03351
```

Test Shapiro-Wilka pokazuje, że są podstawy by odrzucić hipoteze, że dane pochodzą z rozkładu normalnego ponieważ p-value jest mniejsze niż poziom istotności 0.05.

Obliczenia dla temparatury ciała mężczyzn

Weryfikacja hipotezy z założeniem, że wariancja nie jest znana

```r
n_man = length(man_temp$temperatura)

T_man = abs(mean_est_man -mi_0)*sqrt(n_man)/sd_est_man
c_man = qt(1-alfa/2,df = n_man-1)
p_val_man = 2*(1 -pt(T_man, df = n_man-1))
```

Wartość statystyki T = 2.6198952. Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 1.9977297.

p-wartość = 0.010972.

Dla zadanej wartości α hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną.

Wykorzystanie funkcji t.test:

```r
t.test(man_temp$temperatura, mu = mi_0, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  man_temp$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
```

Dla zadanej wartości α hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną.

Za pomocą testu Shapiro-Wilka zweryfikuj hipotezę, że dane pochodzą z rozkładu normalnego:

```r
shapiro.test(man_temp$temperatura)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  man_temp$temperatura
## W = 0.98238, p-value = 0.4818
```

Test Shapiro-Wilka pokazuje, że nie ma podstaw by odrzucić hipoteze, że dane pochodzą z rozkładu normalnego ponieważ p-value jest większe niż poziom istotności 0.05.

Zadanie 2

Treść zadania

W tabeli przedstawionej poniżej zawarto dane dot. liczby samobójstw w Stanach Zjednoczonych w 1970 roku z podziałem na poszczególne miesiące.

Miesiąc	Liczba samobójstw	Liczba dni
Styczeń	1867	31
Luty	1789	28
Marzec	1944	31
Kwiecień	2094	30
Maj	2097	31
Czerwiec	1981	30
Lipiec	1887	31
Sierpień	2024	31
Wrzesień	1928	30
Październik	2032	31
Listopad	1978	30
Grudzień	1859	31

Zweryfikuj czy zamieszczone w niej dane wskazują na sezonową zmienność liczby samobójstw, czy raczej świadczą o stałej intensywności badanego zjawiska. Przyjmij, że w przypadku stałej intensywności liczby samobójstw, liczba samobójstw w danym miesiącu jest proporcjonalna do liczby dni w tym miesiącu.

Rozwiązanie

Zakładamy hipotezę zerową H_0 - dane nie są sezonowe i hipotezę alternatywną H_1 - dane mają charakter sezonowy.

```
ni_i = c(1867, 1789, 1944, 2094, 2097, 1981, 1887, 2024, 1928, 2032, 1978, 1859)
p_i = c(31/365, 28/365, 31/365, 30/365, 31/365, 30/365, 31/365, 31/365,
        30/365, 31/365, 30/365, 31/365)
```

Wykorzystanie funkcji chisq.test:

```
chisq.test(ni_i, p = p_i)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  ni_i
## X-squared = 47.365, df = 11, p-value = 1.852e-06
```

Hipotezę zerową należy odrzucić i przyjąć hipoteze alternatywną mówiacą, że dane mają charakter sezonowy ponieważ wartość p-value jest mnniejsza niż założony poziom istotności wynoszący 0.05. ***

Zadanie 3

Treść zadania

Dla wybranej spółki notowanej na GPW wczytaj dane ze strony bossa.pl

- oblicz wartości procentowych zmian najniższych cen w poszczególnych dniach roku 2019. Wykreśl ich histogram i narysuj funkcję gęstości prawdopodobieństwa rozkładu normalnego o parametrach wyestymowanych na podstawie ich wartości,
- zweryfikuj hipotezę, że procentowe zmiany najniższych cen w poszczególnych dniach roku 2019 mają rozkład normalny.

Rozwiązanie

Załadowanie danych spółki PLAY w 2019 roku i obliczenie wartości procentowych zmian najniższych cen w dniu

```
unzip('mstall.zip', 'PLAY.mst')
df_PLAY = read.csv('PLAY.mst')
names(df_PLAY) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
df_PLAY$date = as.Date.character(df_PLAY$date, format ='%Y%m%d')
df_PLAY = subset(df_PLAY, format.Date(date, "%Y")=="2019")

df_PLAY$low_ch= with(df_PLAY, c(NA, 100*diff(low)/low[-length(low)]))
```

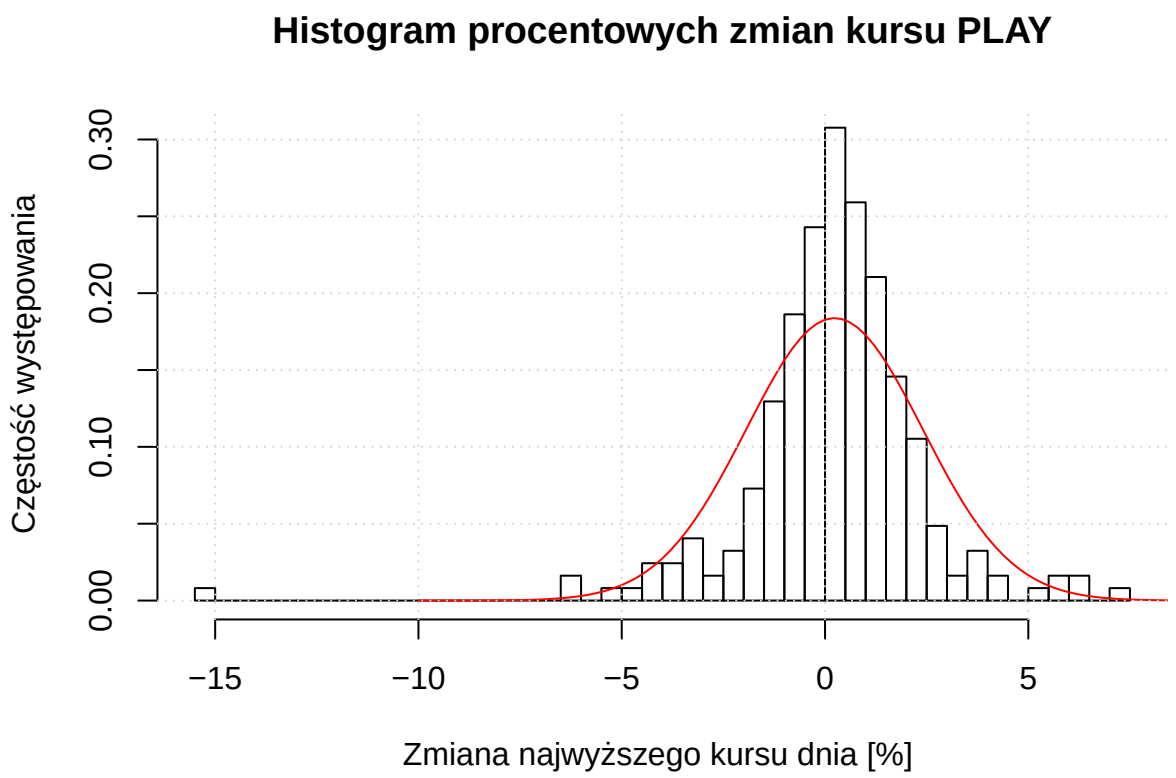
Estymacja wartości średniej i wariancji procentowych zmian najniższych cen w dniu dla spółki PLAY

```
mean_est = mean(df_PLAY$low_ch, na.rm=T)
var_est = var(df_PLAY$low_ch, na.rm=T)
sd_est = sd(df_PLAY$low_ch, na.rm=T)
```

Wartość średnia zmian procentowych wynosi 0.23338. Wartość wariancji zmian procentowych wynosi 4.7125.

Wykreślenie histogramu zmian procentowych i empirycznej funkcji gęstości

```
hist(df_PLAY$low_ch, breaks = 40, prob = T,
xlab = 'Zmiana najwyższego kursu dnia [%] ',
ylab = 'Częstość występowania',
main = 'Histogram procentowych zmian kursu PLAY' )
curve(dnorm(x, mean = mean_est, s = sd_est), add = T, col = 'red', -10, 10)
grid()
```



Przeprowadzenie testu Kolmogorowa-Smirnowa dla średniej równej 0.2333814 i odchylanie standardowego równego 2.170828

```
ks.test(df_PLAY$low_ch, 'pnorm', mean = mean_est, sd = sd_est)
```

```
## Warning in ks.test(df_PLAY$low_ch, "pnorm", mean = mean_est, sd = sd_est):
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_PLAY$low_ch
## D = 0.10088, p-value = 0.01311
## alternative hypothesis: two-sided
```

Na podstawie testu Kolmogorowa_Smirnowa należy odrzucić hipotezę mówiącą, że dane mają rozkład normalny o średniej 0.2333814 i wariancji 2.170828 ponieważ wartość p-value jest mniejsza niż zadany poziom istotności równy 0.05.

Przeprowadzenie testu Shapiro-Wilka

```
shapiro.test(df_PLAY$low_ch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_PLAY$low_ch
## W = 0.88357, p-value = 7.898e-13
```

Na podstawie testu Shapiro-Wilka należy odrzucić hipoteze zerową mówiącą, że dane mają rozkład normalny ponieważ wartość p-value jest mniejsza niż zadany poziom istotności równy 0.05.

Zadanie 4

Treść zadania

W pliku lozyska.txt podane są czasy (w milionach cykli) pracy (do momentu uszkodzenia) łożysk wykonywanych z dwóch różnych materiałów.

- Przeprowadź test braku różnicy między czasami pracy łożysk wykonanych z różnych materiałów, zakładając że czas pracy do momentu uszkodzenia opisuje się rozkładem normalnym.
- Przeprowadź analogiczny test, bez zakładania normalności rozkładów.
- **(dla chętnych)** *Oszacuj prawdopodobieństwo tego, że łożysko wykonane z pierwszego materiału będzie pracowało dłużej niż łożysko wykonane z materiału drugiego.*

Rozwiązanie

Wczytanie danych

```
lozyska = read.csv('lozyska.txt')
```

Test braku różnic miedzy czasami łożysk wykonanych z różnych materiałów przy założeniu rozkładu normalnego

```
t.test(lozyska$X.Typ.I., lozyska$X.Typ.II.)
```

```
##
##  Welch Two Sample t-test
##
## data:  lozyska$X.Typ.I. and lozyska$X.Typ.II.
## t = 2.0723, df = 16.665, p-value = 0.05408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07752643  7.96352643
## sample estimates:
## mean of x mean of y
##    10.693     6.750
```

Na podstawie t-testu przy poziomie istotności równym 0.05 nie ma podstaw do odrzucenia hipotezy zerowej mówiącej, że nie ma różnic w czasie pracy dwóch typów łożysk.

Test braku różnic miedzy czasami łożysk wykonanych z różnych materiałów bez założenia o rozkładzie normalnym

```
wilcox.test(lozyska$X.Typ.I., lozyska$X.Typ.II.)
```

```
##
##  Wilcoxon rank sum test
##
## data:  lozyska$X.Typ.I. and lozyska$X.Typ.II.
## W = 75, p-value = 0.06301
## alternative hypothesis: true location shift is not equal to 0
```

Na podstawie testu Wilcoxona przy poziomie istotności równym 0.05 nie ma podstaw do odrzucenia hipotezy zerowej mówiącej, że nie ma różnic w czasie pracy dwóch typów łożysk.

Zadanie 5

Treść zadania

Korzystając z danych zawartych na stronie pl.fcstats.com zweryfikuj hipotezę o niezależności wyników (zwycięstw, remisów i porażek) gospodarzy od kraju, w którym prowadzone są rozgrywki piłkarskie.

- Dane znajdują się w zakładce Porównanie lig -> Zwycięzcy meczów, w kolumnach (bez znaku [%]):
  - 1 – zwycięstwa gospodarzy, np. dla Bundesligi 145,
  - x – remisy, np. dla Bundesligi 72,
  - 2 – porażki gospodarzy, np. dla Bundesligi 89.
- Testy przeprowadź na podstawie danych dotyczących lig:
  - niemieckiej – Bundesliga,
  - polskiej – Ekstraklasa,
  - angielskiej – Premier League (Liga angielska),
  - hiszpańskiej – Primera Division (Liga hiszpańska).

Rozwiązanie

Zakładamy hipotezę H_0 mówiącą, że dane są niezależne. Zakładamy poziom istotności α równy 0.05.

Obliczenia

```
x_1 = c(125, 95)              #bundesliga
x_2 = c(108,67)               #ekstraklasa
x_3 = c(193,91)               #premier league
x_4 = c(194,91)               #primera division

xx = cbind(x_1, x_2, x_3, x_4)
I = 2
J = 4
n_i =x_1 +x_2+x_3+ x_4
n_j = c(sum(x_1), sum(x_2), sum(x_3), sum(x_4))
N = sum(n_j)
```

Obliczenie wartości statystyki decyzyjnej, progu i p-wartości

```
Tc = 0
for (i in 1:I) {
   for (j in 1:J) {
      Tc = Tc + (N*xx[i,j] -n_i[i]*n_j[j])^2/(N*n_i[i]*n_j[j])
   }
}
alfa = 0.05
```

```
cc = qchisq(1 -alfa, df = (I -1)*(J -1))
p_value_ = 1 - pchisq(Tc, df = (I -1)*(J -1))
```

Wartość statystyki T = 9.2962722. Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 7.8147279. Wartość p-value wynosi 0.0256004. Są więc podstawy by odrzucić hipotezę zerową mówiącą o niezależności danych.

Wykorzystanie funkcji chisq.test:

```
chisq.test(xx)
```

```
##
##  Pearson's Chi-squared test
##
## data:  xx
## X-squared = 9.2963, df = 3, p-value = 0.0256
```

Wartość p-value jest niższa niż zadany poziom istotności 0.05 - są więc podstawy by odrzucić hipotezę o niezależności danych.