ADPS 2021Z — Laboratorium 3 (rozwiązania)

Ola Jaglińska

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 przy poziomie istotności $\alpha = 0.05$ hipotezę, że średnia temperatura jest równa 36.6 °C wobec hipotezy alternatywnej, że średnia temperatura jest inna, przyjmując, że temperatury mają rozkład normalny, a wariancja rozkładu jest nieznana,
- przeprowadź testy normalności dla zarejestrowanych temperatur.

Rozwiązanie

```
temp = read.csv('tempciala.txt')
mtemp = temp[1:65,1]
ktemp = temp[66:130,1]
```

- Wartość średnia i odchylenie standardowe temperatury u mężczyzn

```
m_mtemp = mean(mtemp)
sd_mtemp = sd(mtemp)
```

Średnia wartość temperatury mężczyzn wynosi: 36.7262, a odchylenie standardowe: 0.3882.

- zweryfikuj przy poziomie istotności $\alpha = 0.05$ hipotezę, że średnia temperatura jest równa 36.6 °C wobec hipotezy alternatywnej, że średnia temperatura jest inna, przyjmując, że temperatury mają rozkład normalny, a wariancja rozkładu jest nieznana,

```
z1_mi_0 = 36.6; alfa = 0.05; z1_n = 65
mT = abs(m_mtemp - z1_mi_0)*sqrt(z1_n)/sd_mtemp
mc = qnorm(1 - alfa/2)
mp_val = 2*(1 - pnorm(mT))
```

Wartość statystyki dla mężczyzn T = 2.6199.

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 1.96.

Na podstawie powyższych wyników wnioskuję, że p-wartość p-val = 0.0088 jest mniejsza od alfy i mogę odrzucić hipotezę zerową. Prawdopodobnie jest jakiś błąd w obliczeniach p-value, ponieważ wychodzi zaskakująco niska.

- Testy normalności dla zarejestrowanych temperatur.

Test Kołmogorowa-Smirnowa

```
ks.test(mtemp, 'pnorm', m_mtemp, sd_mtemp)
```

```
## Warning in ks.test(mtemp, "pnorm", m_mtemp, sd_mtemp): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  mtemp
## D = 0.088528, p-value = 0.6883
## alternative hypothesis: two-sided
```

P-wartośc wychodzi duża -> dane mają rozkład normalny

Test Shapiro-Wilka:

```
shapiro.test(mtemp)
```

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

P-wartośc wychodzi duża -> dane mają rozkład normalny.

Test Anscombe-Glynn na kurtozę:

```
library(normtest)
kurtosis.norm.test(mtemp)
```

```
##
##  Kurtosis test for normality
##
## data:  mtemp
## T = 2.6135, p-value = 0.488
```

P-wartośc wychodzi duża -> dane mają rozkład normalny.

Test Jarque-Bera

```
library(moments)
jarque.test(mtemp)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  mtemp
## JB = 1.197, p-value = 0.5496
## alternative hypothesis: greater
```

P-wartośc wychodzi duża -> dane mają rozkład normalny.

Test D'Agostino

```
library(moments)
agostino.test(mtemp)
```

```
##
##  D'Agostino skewness test
##
## data:  mtemp
## skew = -0.27047, z = -0.96033, p-value = 0.3369
## alternative hypothesis: data have a skewness
```

P-wartośc wychodzi duża -> dane mają rozkład normalny.

- Wartość średnia i odchylenie standardowe temperatury u kobiet

```
m_ktemp = mean(ktemp)
sd_ktemp = sd(ktemp)
```

Średnia wartość temperatury kobiet wynosi: 36.8892, a odchylenie standardowe: 0.4127.

- zweryfikuj przy poziomie istotności $\alpha = 0.05$ hipotezę, że średnia temperatura jest równa 36.6 °C wobec hipotezy alternatywnej, że średnia temperatura jest inna, przyjmując, że temperatury mają rozkład normalny, a wariancja rozkładu jest nieznana,

```
z1_mi_0 = 36.6; alfa = 0.05; z1_n = 65
kT = abs(m_ktemp - z1_mi_0)*sqrt(z1_n)/sd_ktemp
kc = qnorm(1 - alfa/2)
kp_val = 2*(1 - pnorm(kT))
```

Wartość statystyki dla kobiet kT = 5.6497.

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 1.96.

Na podstawie powyższych wyników wnioskuję, że p-wartość p-val = 1.6068566×10^{-8} jest większa od alfy i nie mogę odrzucić hipotezy zerowej.

- przeprowadź testy normalności dla zarejestrowanych temperatur.

Test Kołmogorowa-Smirnowa

```
ks.test(ktemp, 'pnorm', m_ktemp, sd_ktemp)
```

```
## Warning in ks.test(ktemp, "pnorm", m_ktemp, sd_ktemp): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ktemp
## D = 0.12018, p-value = 0.3049
## alternative hypothesis: two-sided
```

P-wartośc wychodzi duża -> dane mają rozkład normalny

Test Shapiro-Wilka:

```
shapiro.test(ktemp)
```

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

P-wartośc wychodzi mała -> dane nie mają rozkładu normalnego.

Test Anscombe-Glynn na kurtozę:

```r
library(normtest)
kurtosis.norm.test(ktemp)
```

```
##
##  Kurtosis test for normality
##
## data:  ktemp
## T = 4.4251, p-value = 0.013
```

P-wartośc wychodzi mała -> dane nie mają rozkładu normalnego.

Test Jarque-Bera

```r
library(moments)
jarque.test(ktemp)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  ktemp
## JB = 5.5021, p-value = 0.06386
## alternative hypothesis: greater
```

P-wartośc wychodzi większa niż alfa -> dane mają rozkład normalny.

Test D'Agostino

```r
library(moments)
agostino.test(ktemp)
```

```
##
##  D'Agostino skewness test
##
## data:  ktemp
## skew = 0.011492, z = 0.041338, p-value = 0.967
## alternative hypothesis: data have a skewness
```

P-wartośc wychodzi duża -> dane mają rozkład normalny.

Nie wiem skąd pojawiła się ta rozbieżność w wynikach testów.

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

Miesiąc	Liczba samobójstw	Liczba dni
Czerwiec	1981	30
Lipiec	1887	31
Sierpień	2024	31
Wrzesień	1928	30
Październik	2032	31
Listopad	1978	30
Grudzień	1859	31

Zweryfikuj przy poziomie istotności $\alpha = 0.05$ 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

```
z2_ni_i = c(1867, 1789, 1944, 2094, 2097, 1981, 1887, 2024, 1928, 2032, 1978, 1859)
z2_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, 31/3
z2_n = sum(z2_ni_i)
z2_T = sum((z2_ni_i - z2_n*z2_p_i)^2 / (z2_n*z2_p_i))
z2_r = 12
alfa = 0.05
z2_c = qchisq(1 - alfa, z2_r - 1)
z2_p_val = 1 - pchisq(z2_T, z2_r - 1)
```

Wartość statystyki T = 47.3653.

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 19.6751.

p-wartość = 0.

```
chisq.test(z2_ni_i, p = z2_p_i)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  z2_ni_i
## X-squared = 47.365, df = 11, p-value = 1.852e-06
```

Bardzo mała p-wartość, dużo niższa od $\alpha = 0.05$, wskazuje na odrzucenie hipotezy zerowej.

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 2021, wykreśl ich histogram i narysuj funkcję gęstości prawdopodobieństwa rozkładu normalnego o parametrach wyestymowanych na podstawie ich wartości,

- stosując rózne testy omawiane w przykładach zweryfikuj przy poziomie istotności $\alpha = 0.05$ hipotezę, że procentowe zmiany najniższych cen w poszczególnych dniach roku 2021 mają rozkład normalny.

Rozwiązanie

```
  if(!file.exists('mstall.zip')) {
    download.file('https://info.bossa.pl/pub/metastock/mstock/mstall.zip','mstall.zip')
  }
```

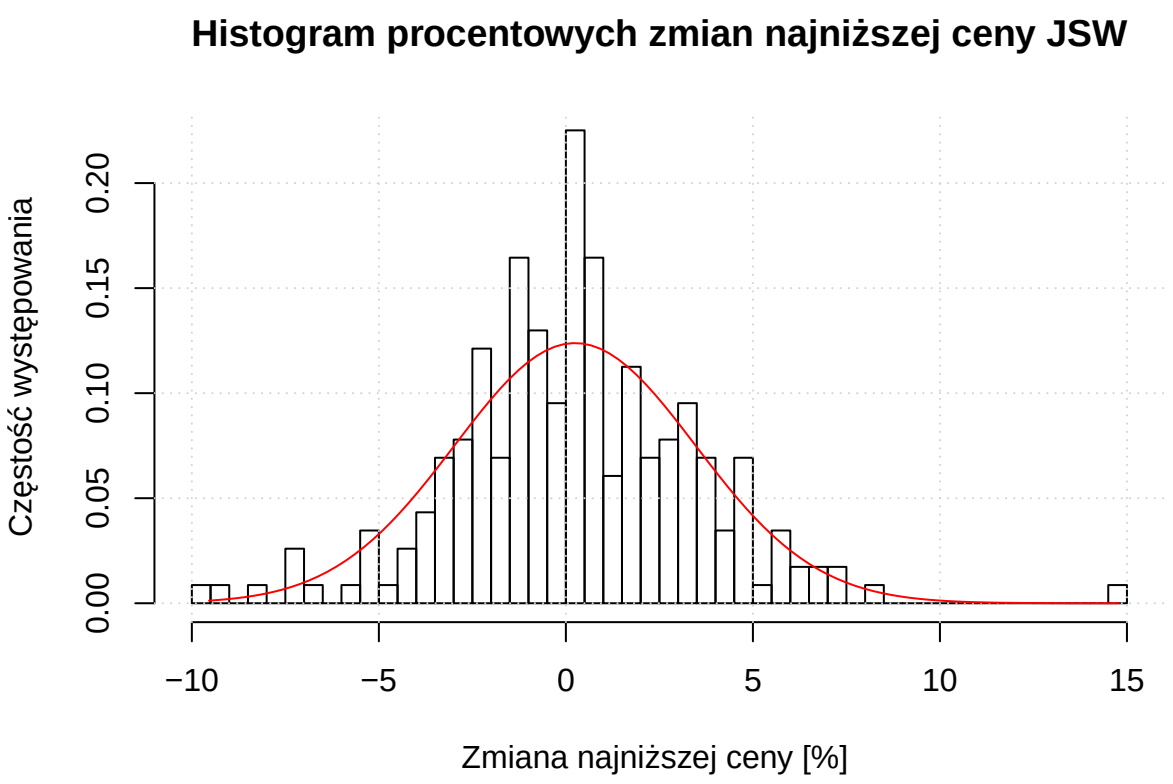
```
unzip('mstall.zip', 'JSW.mst')
df_JSW = read.csv('JSW.mst')
```

- oblicz wartości procentowych zmian najniższych cen w poszczególnych dniach roku 2021, wykreśl ich histogram i narysuj funkcję gęstości prawdopodobieństwa rozkładu normalnego o parametrach wyestymowanych na podstawie ich wartości

```
names(df_JSW) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
df_JSW$date = as.Date.character(df_JSW$date, format ='%Y%m%d')
df_JSW = df_JSW[which(df_JSW$date >= '2021-01-01' & df_JSW$date <= '2021-12-02'),]
df_JSW$low_ch = with(df_JSW, c(NA, 100*diff(low)/low[-length(low)]))
m_jsw = mean(df_JSW$low_ch, na.rm = T)
v_jsw = var(df_JSW$low_ch, na.rm = T)
s_jsw = sd(df_JSW$low_ch, na.rm = T)
```

Histogram z funkcją gęstości prawdopodobieństwa:

```
hist(df_JSW$low_ch, breaks = 50, prob = T,
xlab = 'Zmiana najniższej ceny [%] ',
ylab = 'Częstość występowania',
main = paste('Histogram procentowych zmian najniższej ceny JSW')  )
grid()
min_c = min(df_JSW$low_ch, na.rm = T)
max_c = max(df_JSW$low_ch, na.rm = T)
curve(dnorm(x, mean = m_jsw, sd = s_jsw), add = T, col = 'red', from = min_c, to = max_c)
```



- stosując rózne testy omawiane w przykładach zweryfikuj przy poziomie istotności $\alpha = 0.05$ hipotezę, że procentowe zmiany najniższych cen w poszczególnych dniach roku 2021 mają rozkład normalny.

Test Kołmogorowa-Smirnowa:

```
ks.test(df_JSW$low_ch, 'pnorm', mean = m_jsw, sd = s_jsw)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_JSW$low_ch
## D = 0.061363, p-value = 0.3493
## alternative hypothesis: two-sided
```

P-wartość wyższa od alfy -> nie mogę odrzucić hipotezy zerowej. Zmiany najniższych cen mają rozkład normalny.

Test Shapiro-Wilka:

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

```
##
##  Shapiro-Wilk normality test
##
## data:  df_JSW$low_ch
## W = 0.98136, p-value = 0.003885
```

P-wartość niższa od alfy -> mogę odrzucić hipotezę zerową. Zmiany najniższych cen nie mają rozkładu normalnego.

Test D'Agostino:

```
library(moments)
agostino.test(df_JSW$low_ch)
```

```
##
##  D'Agostino skewness test
##
## data:  df_JSW$low_ch
## skew = 0.19001, z = 1.20483, p-value = 0.2283
## alternative hypothesis: data have a skewness
```

P-wartość wyższa od alfy -> nie mogę odrzucić hipotezy zerowej. Zmiany najniższych cen mają rozkład normalny.

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, bez zakładania równości wariancji. Przyjmij poziom istotności $\alpha = 0.05$.
- 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

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

- Przeprowadź test braku różnicy między czasami pracy łożysk wykonanych z różnych materiałów, zakładając że czas pracy do momentu uszkodzenia opisuje się rozkładem normalnym, bez zakładania równości wariancji. Przyjmij poziom istotności $\alpha = 0.05$. ‘

```
names(lozyska) = c('X.Typ.I', 'X.Typ.II')
z4_n = 10
typ.1 = lozyska$X.Typ.I
typ.2 = lozyska$X.Typ.II
m_typ1 = mean(typ.1); m_typ2 = mean(typ.2)
s_typ1 = var(typ.1); s_typ2 = var(typ.2)
z4_s = s_typ1/z4_n + s_typ2/z4_n
z4_d = (z4_s^2)/(((s_typ1/z4_n)^2)/(z4_n-1) + ((s_typ2/z4_n)^2)/(z4_n-1))
alfa = 0.05
z4_c = qt(1 - alfa/2, z4_d)
z4_T = abs(m_typ1 -m_typ2)/sqrt(z4_s)
z4_p_val = 2*(1 - pt(z4_T, z4_d))
```

Wartość statystyki T = 2.0723.

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 2.1131.

p-wartość = 0.0541.

```
t.test(typ.1, typ.2)
```

```
##
##  Welch Two Sample t-test
##
## data:  typ.1 and typ.2
## 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
```

P-wartośc jest wyższa od alfy -> nie odrzucam hipotezy zerowej.

- Przeprowadź analogiczny test, bez zakładania normalności rozkładów.

```
wilcox.test(typ.1, typ.2)
```

```
##
##  Wilcoxon rank sum test
##
## data:  typ.1 and typ.2
## W = 75, p-value = 0.06301
## alternative hypothesis: true location shift is not equal to 0
```

P-wartośc jest wyższa od alfy -> nie odrzucam hipotezy zerowej.

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. Przyjmij poziom istotności $\alpha = 0.05$.

- Testy przeprowadź na podstawie danych dotyczących lig:
  - niemieckiej – Bundesliga (Liga niemiecka),
  - polskiej – Ekstraklasa (Liga polska),
  - angielskiej – Premier League (Liga angielska),
  - hiszpańskiej – Primera Division (Liga hiszpańska).
- Dane znajdują się w zakładce Porównanie lig -> Zwycięzcy meczów, w kolumnach (bez znaku [%]):
  - 1 – zwycięstwa gospodarzy, np. dla ligi niemieckiej (Bundesliga) 125,
  - x – remisy, np. dla ligi niemieckiej 86,
  - 2 – porażki gospodarzy, np. dla ligi niemieckiej 95.

Rozwiązanie

```
Bundesliga = c(125, 86, 95)
Ekstraklasa = c(108, 65, 67)
Premier_League = c(193, 96, 91)
Primera_Division = c(194, 95, 91)
ligi = cbind(Bundesliga, Ekstraklasa, Premier_League, Primera_Division)
z5_I = 3
z5_J = 4
z5_n_i = Bundesliga + Ekstraklasa + Premier_League + Primera_Division
z5_n_j = c(sum(Bundesliga), sum(Ekstraklasa), sum(Premier_League), sum(Primera_Division))
z5_N = sum(z5_n_j)
```

Obliczenie wartości statystyki decyzyjnej, progu i p-wartości:

```
z5_T = 0
for (z5_i in 1:z5_I) {
   for (z5_j in 1:z5_J) {
      z5_T = z5_T + (z5_N*ligi[z5_i,z5_j] - z5_n_i[z5_i]*z5_n_j[z5_j])^2/(z5_N*z5_n_i[z5_i]*z5_n_j[z5_j
   }
}
alfa = 0.05
z5_c = qchisq(1 - alfa, df = (z5_I - 1)*(z5_J - 1))
z5_p_val = 1 - pchisq(z5_T, df = (z5_I - 1)*(z5_J - 1))
```

Wartość statystyki T = 10.3254.

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ wynosi c = 12.5916.

p-wartość = 0.1116.

Sprawdzam swoje wyniki przy użyciu funkcji chisq.test:

```
chisq.test(ligi)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ligi
## X-squared = 10.325, df = 6, p-value = 0.1116
```

P-wartośc zgadza się z moimi wyliczeniami.

Wniosek: P-wartość jest wyższa od alfy, a statystyka T = 10.3254 jest mniejsza od wartości krytycznej c = 12.5916, nie mam podstaw do odrzucenia hipotezy zerowej o niezależności cech.
