ADPS 2025L — Laboratorium 3 (rozwiązania)

Dariusz Kopka

Zadanie 1 (1 pkt)

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.

Rozwiązanie

Osobno dla mężczyzn i kobiet: wyestymuj wartość średnią i odchylenie standardowe temperatury

```
require(readr, quietly = TRUE)
d <- read_csv('tempciala.txt', show_col_types = FALSE)
men <- d$temperatura[d$płeć == 1]
women <- d$temperatura[d$płeć == 2]

men_avg <- mean(men)
women_avg <- mean(women)

men_sd <- sd(men)
women_sd <- sd(women)
```

Dla grupy kobiet wartość średnia temperatury wynosi 36.8892308°C przy odchyleniu standardowym 0.4127359.

Odpowiednio dla grupy mężczyzn średnia temperatura wynsi 36.7261538°C, a odchylenie standardowe 0.3882158.

Osobno dla mężczyzn i kobiet: 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.

```
mu_0 <- 36.6
alfa <- 0.05
gamma <- 1 - alfa

# Test t-Studenta dla jednej próby dla mężczyzn i kobiet osobno
t.test(men,   mu = mu_0, alternative = "two.sided", conf.level = gamma) -> men_test
t.test(women, mu = mu_0, alternative = "two.sided", conf.level = gamma) -> women_test
men_test
```

```
##
##  One Sample t-test
##
## data:  men
## 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
```

```
women_test
```

```
##
##  One Sample t-test
##
## data:  women
## 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
```

Przeprowadzono test t-Studenta dla hipotezy zerowej $H_0 : \mu = 36.6°\text{C}$ oraz hipotezy alternatywnej $H_1 : \mu \neq 36.6°\text{C}$ przy poziomie ufności $\gamma = 1 - \alpha = 0.95$.

Test wykazał p-wartość dla grupy mężczyzn na poziomie p = 0.010972, a dla grupy kobiet na poziomie p $= 3.9852716 \times 10^{-7}$ co na poziomie istotności $\alpha = 0.05$ skutkuje odrzuceniem hipotezy zerowej H_0 dla obu badanych grup.

Należy zatem przyjąć, że średnia temperatura ciała zarówno dla kobiet jak i mężczyzn **nie jest równa** 36.6°C (hipoteza alternatywna H_1 jest prawdziwa).

Zadanie 2 (1 pkt)

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 przy poziomie istotności $\alpha = 0.05$ czy zamieszczone w niej dane świadczą o stałej intensywności badanego zjawiska, czy raczej wskazują na sezonową zmienność liczby samobójstw. Przyjmij, że w przypadku stałej intensywności liczby samobójstw, liczba samobójstw w danym miesiącu jest proporcjonalna do liczby dni w tym miesiącu.

Rozwiązanie

Jako hipotezę zerową H_0 przyjmuję, że liczba samobójstw jest proporcjonalna do liczby dni w miesiącu. Oznacza to, że rozkład prawdopodobieństwa wystąpienia samobójstwa w ciągu roku jest proporcjonalny do liczby dni w danym miesiącu. Na przykład dla stycznia, który ma 31 dni, prawdopodobieństwo wynosi $\frac{31}{365} \approx 0.0849315$.

```
dane <- data.frame(
  miesiac = seq(1:12),
  liczba_dni = c(31, 28, 31, 30, 31, 30,
                 31, 31, 30, 31, 30, 31),
  samobojstwa = c(1867, 1789, 1944, 2094, 2097, 1981,
                  1887, 2024, 1928, 2032, 1978, 1859)
)

suma_samobojstw <- sum(dane$samobojstwa)
dane$czesc_roku <- dane$liczba_dni / sum(dane$liczba_dni)

chisq.test(x = dane$samobojstwa, p = dane$czesc_roku) -> samobojstwa_test
samobojstwa_test
```

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

Test chi-kwadrat wykazał, że hipotezę zerową H_0 należy odrzucić, ponieważ p-wartość $p = 1.8520112 \times 10^{-6}$ jest mniejsza niż założony poziom istotności $\alpha = 0.05$.

Oznacza to, że hipotezę zerową H_0 należy odrzucić oraz że liczba samobójstw w badanych danych wykazuje się sezonowością.

Zadanie 3 (1 pkt)

Treść zadania

Dla wybranej spółki notowanej na GPW wczytaj dane ze strony stooq.pl, a następnie

- oblicz wartości procentowych zmian najniższych cen w poszczególnych dniach w ciągu ostatniego roku, 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 w ciągu ostatniego roku mają rozkład normalny.

Rozwiązanie

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

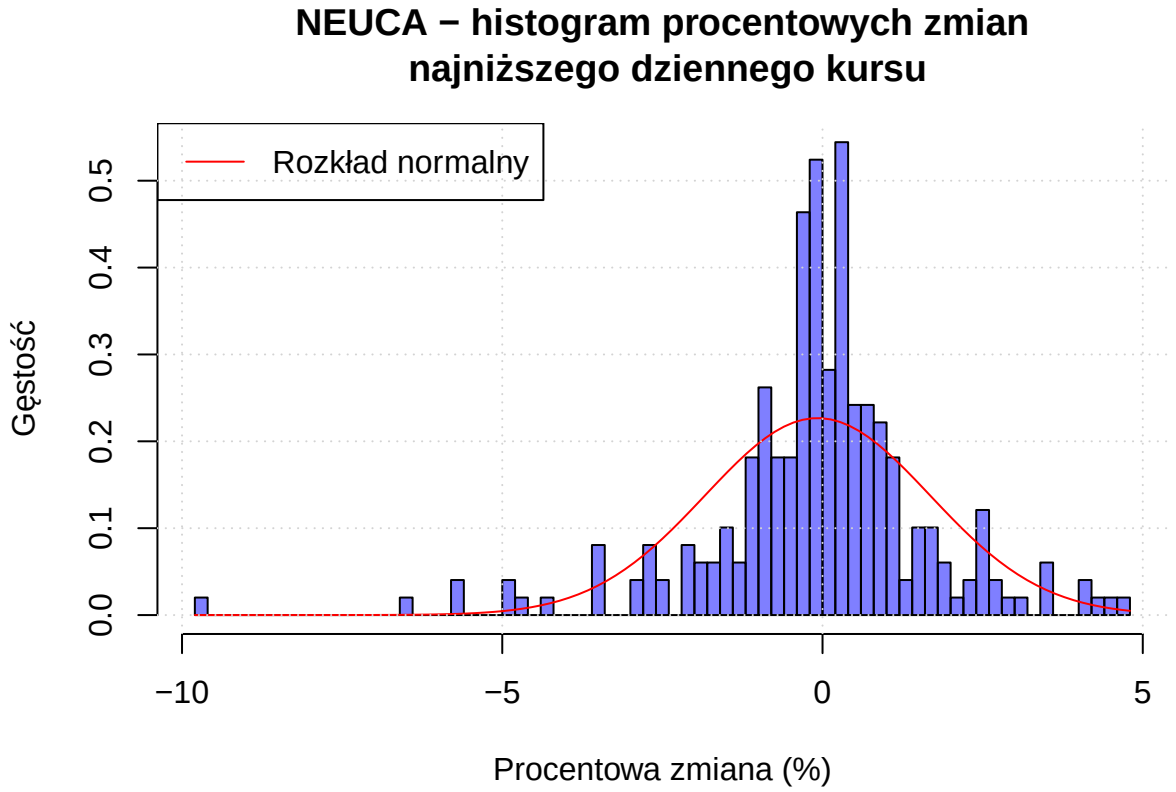
```
# Cały ten fragment to copy-paste ze sprawozdania do Lab 2.
ticker = list("NEUCA" = 'neu')
# Wykorzystuję tu funkcję, którą napisałem w sprawozdaniu do pierwszego laboratorium
akcje <- get_stock(ticker[[1]], start_date = "2024-04-05", end_date = "2025-04-05")
akcje$Date <- as.Date(akcje$Date)
akcje$Low <- as.numeric(akcje$Low)

akcje$LowChange <- c(NA, 100 * diff(akcje$Low) / head(akcje$Low, -1))
dane <- na.omit(akcje$LowChange)

# Estymacja średniej, wariancji i odchylenia standardowego
sd_mean <- mean(dane)
sd_var <- var(dane)
sd_sd <- sd(dane)

hist(akcje$LowChange,
     breaks = 60,
     col = rgb(0, 0, 1, 0.5),
     border = "black",
     main = paste0(names(ticker[1]), " - histogram procentowych zmian",
                                     "\n najniższego dziennego kursu"),
     ylab = "Gęstość",
     xlab = "Procentowa zmiana (%)",
```

```
     probability = TRUE)
curve(dnorm(x, mean = sd_mean, sd = sd_sd),
      col = "red", lwd = 1, add = TRUE)
legend("topleft",
       legend = "Rozkład normalny",
       col =  "red",
       lty = 1
)
grid()
```



Stosując różne testy omawiane w przykładach zweryfikuj przy poziomie istotności $\alpha = 0.05$ hipotezę, że procentowe zmiany najniższych cen w poszczególnych dniach w ciągu ostatniego roku mają rozkład normalny

```
x <- as.vector(na.omit(akcje$LowChange))

ks.test(x, 'pnorm',  alternative = "two.sided", mean = mean(x), sd = sd(x)) -> ks_test
```

```
## Warning in ks.test.default(x, "pnorm", alternative = "two.sided", mean =
## mean(x), : ties should not be present for the one-sample Kolmogorov-Smirnov
## test
```

```
ks_test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.11828, p-value = 0.001939
## alternative hypothesis: two-sided
```

Test zgodności Kołmogorowa-Smirnowa wykazał wartość statystyki testowej $D = 0.1182786$, co oznacza, że największa bezwzględna różnica między dystrybuantą empiryczną a dystrybuantą rozkładu normalnego wynosi 11.8278644%. p-wartość wynosi $p = 0.0019386$ i jest poniżej progu istotności $\alpha = 0.05$, co stanowi podstawę do odrzucenia hipotezy zerowej H_0.

Oznacza to, że rozkład danych istotnie różni się od rozkładu normalnego.

```
shapiro.test(x) -> shapiro_test
shapiro_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.91123, p-value = 5.737e-11
```

Test Shapiro-Wilka jest najmocniejszym testem sprawdzającym normlaność. Otrzymana statystyka testowa $W = 0.9112319$ pokazuje istotne odchylenie danych od rozkładu normalnego. Otrzymana p-wartość $p = 5.7370246 \times 10^{-11}$ jest znacznie poniżej przyjętego progu istotności $\alpha = 0.05$ i pozwala odrzucić hipotezę zerową jakoby rozkład danych procentowych zmian najniższej dziennej ceny akcji był rozkladem normalnym.

```
library(moments)
anscombe.test(x, alternative = "two.sided") -> anscombe_test
anscombe_test
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  x
## kurt = 7.6998, z = 5.7804, p-value = 7.452e-09
## alternative hypothesis: kurtosis is not equal to 3
```

Test kurtozy Anscombe-Glynn'a mówi o tym jak 'spiczasty' jest rozkład empiryczny. Dla rozkładu normalnego kurtoza jest równa 3. Wartości powyżej 3 mówią o większej spiczastości rozkładu (leptokurtozie) i grubszych ogonach. Wartości poniżej 3 świadczą o mniejszym skoncentrowaniu wokół wartości średniej i mniejszej ilości wartości ekstremalnych (platykurtyzie). Hipotezą zerową dla testu Anscombe-Glynn'a jest kurtoza równa 3. Hipotezą alternatywną jest tutaj kurtoza różna od 3.

Dla badanego rozkładu kurtoza wskazana przez test Anscombe-Glynn'a wykazała wartość `Kurt` =7.6997792 co świadczy o bardzo dużym skoncentrowaniu wokół wartości średniej oraz dużej ilości wartości ekstremalnych. Otrzymana p-wartość $p = 7.451517 \times 10^{-9}$ jest również znacznie poniżej progu istotności ustwaionego na $\alpha = 0.05$. Hipotezę zerową należy odrzucić.

```r
agostino.test(x) -> agostino_test
agostino_test
```

```
##
##  D'Agostino skewness test
##
## data:  x
## skew = -1.0177, z = -5.7042, p-value = 1.169e-08
## alternative hypothesis: data have a skewness
```

Test skośności D'Agostino sprawdza czy rozkład danych jest symetryczny. Otrzymana w teście wartość `skew` = -1.0176501 świadczy o lewostronnej skośności rozkładu (bo wartość jest ujemna). p-wartość $p = 1.1688262 \times 10^{-8}$ jest znacznie poniżej progu istotności $\alpha = 0.05$.

Daje to kolejną przesłankę do twierdzenia, że dane nie są 'normalne'.

```r
jarque.test(x) -> jarque_test
jarque_test
```

```
##
##  Jarque-Bera Normality Test
##
## data:  x
## JB = 271.05, p-value < 2.2e-16
## alternative hypothesis: greater
```

Test normalności Jarque-Bera sprawdza zarówno skośność jak i kurtozę, oczekując, że skośność = 0 i kurtoza = 3.

Otrzymana p-wartość $p \approx 0$ nakazuje odrzucić hipotezę zerową, że dane należą do rozkładu normalnego.

Zadanie 4 (1 pkt)

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

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 $\alpha = 0.05$

```
lozyska <-  read.csv('lozyska.txt')
typI <- lozyska$X.Typ.I.
typII <- lozyska$X.Typ.II.

# var.equal = FALSE oznacza, że nie zakładamy, że wariancje nie są równe.
t.test(typI, typII, alternative = "two.sided", var.equal = FALSE) -> welch_test
welch_test
```

```
##
##  Welch Two Sample t-test
##
## data:  typI and typII
## 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
```

Ponieważ test t-Studenta dla dwóch prób niezależnych, nazywany testem Welcha, wykazał p-wartość na poziomie 0.0540794, brakuje podstaw do odrzucenia hipotezy zerowej, choć wartość znajduje się bardzo blisko granicy istotności.

Hipoteza zerowa H_0 dla testu Welch'a mówi o tym, że różnica średnich obu badanych próbek *nie jest* istotana statystycznie. Analogicznie hipoteza alternatywna H_1 mówi, że różnica średnich *jest* istotna statystycznie.

Zatem nie stwierdzono istotnych statystycznie różnic średnich czasów pracy łożysk typu pierwszego i typu drugiego (H_0).

Przeprowadź analogiczny test, bez zakładania normalności rozkładów.

```
wilcox.test(typI, typII, alternative = "two.sided") -> wilcox_test
wilcox_test
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  typI and typII
## W = 75, p-value = 0.06301
## alternative hypothesis: true location shift is not equal to 0
```

Hipotezy dla testu Wilcoxona: H_0 - czasy pracy obu typów łożysk są identyczne, H_1 - czasy różnią się

Test wykazał p-wartość $p = 0.0630128$ co na poziomie istotności $\alpha = 0.05$ nie pozwala odrzucić hipotezy zerowej. W związku z tym, nie stwierdzono istotnej statystycznie różnicy pomiędzy czasami pracy łożysk z materiałów I i II na poziomie istotności $\alpha = 0.05$.

(dla chętnych) Oszacuj prawdopodobieństwo tego, że łożysko wykonane z pierwszego materiału będzie pracowało dłużej niż łożysko wykonane z materiału drugiego.

Za Wikipedia:

> Metoda Monte Carlo - metoda stosowana do modelowania matematycznego procesów zbyt złożonych (obliczania całek, łańcuchów procesów statystycznych), aby można było przewidzieć ich wyniki za pomocą podejścia analitycznego. Istotną rolę w tej metodzie odgrywa losowanie (wybór przypadkowy) wielkości charakteryzujących proces, przy czym losowanie dokonywane jest zgodnie z rozkładem, który musi być znany.

Na studiach (Elektronika i Telekomunikacja) używałem tej metody do analizy obwodów elektrycznych, a w szczególności wpływu tolerancji parametrów komponentów pasywnych na zachowanie układu (np. seria rezystorów o tolerancji $\pm$ 5%).

```
n_sim <- 25000
probki_typ1 <- sample(typI, size = n_sim, replace = TRUE)
probki_typ2 <- sample(typII, size = n_sim, replace = TRUE)
prawd_1_zyje_dluzej_niz_2 <- mean(probki_typ1 > probki_typ2)
```

Oszacowane metodą Monte Carlo prawdopodobieństwo, że łożysko typu 1 będzie pracowało dłużej niż to typu 2 wynosi 75.144%.

Zadanie 5 (1 pkt)

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 – LaLiga (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

```r
liga <- data.frame(
    zwyciestwa = c(125, 193, 108, 194),
    remisy = c(86, 96, 65, 95),
    porazki = c(95, 91, 67, 91)
)
rownames(liga) <- c("Niemcy", "Anglia", "Polska", "Hiszpania")

chisq.test(liga) -> chi_test
chi_test
```

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

Chi-squared test `chisq.test()`, kiedy dostaje jako argument wejściowy tak skonstruowaną ramkę danych przyjmuje, że jest to dwuwymiarowa tabela kontyngencji. Za manualem:

> If x is a matrix with at least two rows and columns, it is taken as a two-dimensional contingency table: he entries of x must be non-negative integers

Został wykononany test χ^2, gdzie:

H_0 - rozkład wyników (1, X, 2) nie zależy od ligi

H_1 - rozkład wyników zależy od ligi

Otrzymana p-wartość $p = 0.111602$ jest powyżej przyjętego poziomu istotności $\alpha = 0.05$, co nie pozwala odrzucić hipotezy zerowej, że wyniki nie zależą od ligi.

Oznacza to, że nie stwierdzono istotnych statystycznie różnic w rozkładzie wyników pomiędzy analizowanymi ligami krajowymi – rozkład wyników może być uznany za niezależny od kraju.
