Analiza Danych Podstawy Statystyczne (ADPS)

Laboratorium 3

Przykład – test wartości średniej

- Wygeneruj 30 liczb będących próbą z rozkładu N(1,4):

  n = 30; mi = 1; sigma = 2

  x = rnorm(n, mean = mi, sd = sigma)

- Przy poziomie istotności α = 0.05 zweryfikuj hipotezę zerową:

$$H_0: \ \mu = 1,$$

$$H_1: \ \mu \neq 1,$$

- przy założeniu, że:
  - wariancja rozkładu jest znana,
  - wariancja rozkładu jest nieznana.

Przykład – test wartości średniej

- Znana wariancja:

  mi_0 = 1; alfa = 0.05

  Z = abs(mean(x) - mi_0)*sqrt(n)/sigma

  c = qnorm(1-alfa/2)

  p_val = 2*(1 - pnorm(Z))

- Wykorzystanie funkcji pakietu R:

  z.test(x, mu = mi_0, sigma, alternative = "two.sided")

* Funkcja z.test znajduje się w pakiecie *TeachingDemos*.

Przykład – test wartości średniej

- Nieznana wariancja:

  T = abs(mean(x) - mi_0)*sqrt(n)/sd(x)

  c = qt(1-alfa/2,df = n-1)

  p_val = 2*(1 - pt(T, df = n-1))

- Wykorzystanie funkcji pakietu R:

  t.test(x, mu = mi_0, alternative = "two.sided")

- Powtórz testy w przypadku znanej i nieznanej wariancji dla innych wartości mi_0 = 0, 1.5, 3.

Przykład – test zgodności Pearsona

- Zasymuluj 50 rzutów kostką i sprawdź, czy kostka jest symetryczna.

```
n = 50; x = sample(1:6, n, replace = T)
ni_i = as.data.frame(table(factor(x, levels = 1:6)))$Freq
p_i = rep(1/6, 6)
T = sum((ni_i - n*p_i)^2 / (n*p_i))
r = 6
alfa = 0.05
c = qchisq(1 - alfa, r - 1)
p_val = (1 - pchisq(T, r - 1))
```

- Wykorzystanie funkcji pakietu R:

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

Przykład – test Kołmogorowa-Smirnowa

- Wygeneruj 50 liczb będących próbą z rozkładu jednostajnego w przedziale od 2 do 4:

  ```
  n = 50; a = 2; b = 4
  x = runif(n, min = a, max = b)
  ```

- Obejrzyj wartości próbek i ich histogram.
- Przeprowadź test Kołmogorowa-Smirnowa dla różnych założeń dotyczących krańców przedziałów:

  ```
  ks.test(x, 'punif', min = a, max = b)
  ks.test(x, 'punif', a - 0.5, b - 0.5)
  ks.test(x, 'punif', a - 0.5, b + 0.5)
  ```

Przykład – test Kołmogorowa-Smirnowa

- Zweryfikuj hipotezę, że wygenerowane dane pochodzą z rozkładu normalnego o parametrach będących wartością średnia i wariancją rozkładu jednostajnego:

```
ks.test(x, 'pnorm', mean = (a + b)/2, sd = sqrt((b - a)^2/12))
```

- Sprawdź wynik po zwiększeniu liczby próbek:

```
x = runif(500, min = a, max = b)
ks.test(x, 'pnorm', mean = (a + b)/2, sd = sqrt((b - a)^2/12))
```

- Zweryfikuj czy dane z rozkładu normalnego dopasowują się do rozkładu jednostajnego:

```
n = 100; mi = 1; sigma = 2
x = rnorm(n, mean = mi, sd = sigma)
ks.test(x, 'punif', min = mi-sigma, max = mi+sigma)
```

Przykład – test normalności

- Wygeneruj 100 liczb będących próbą losową z rozkładu N(1,4):

```
n = 100; mi = 1; sigma = 2
x = rnorm(n, mean = mi, sd = sigma)
```

- Korzystając z testu Kołmogorowa-Smirnowa zweryfikuj hipotezę, że dane pochodzą z rozkładu N(1,4) lub N(0,9):

```
ks.test(x, 'pnorm', mean = 1, sd = 2)
ks.test(x, 'pnorm', mean = 0, sd = 3)
```

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

```
shapiro.test(x)
```

Przykład – test normalności

- Wygeneruj 100 liczb będących próbą losową z rozkładu jednostajnego w przedziale od 2 do 4:

```
n = 100; a = 2; b = 4
x = runif(n, min = a, max = b)
```

- Zweryfikuj hipotezę, że dane pochodzą z rozkładu normalnego korzystając z testu Shapiro-Wilka:

```
shapiro.test(x)
```

- Powtórz weryfikację tej hipotezy korzystając z testu Kołmogorowa-Smirnowa:

```
ks.test(x, 'pnorm', mean = (a + b)/2, sd = sqrt((b - a)^2/12))
```

Przykład – porównywanie średnich

- Wygeneruj dwie próby z rozkładów normalnych N(0,1) oraz N(1,1) o licznościach 9 i 12:

```
n_1 = 9; mi_1 = 0; sigma_1 = 1
n_2 = 12; mi_2 = 1; sigma_2 = 1
x_1 = rnorm(n_1, mean = mi_1, sd = sigma_1)
x_2 = rnorm(n_2, mean = mi_2, sd = sigma_2)
```

Przykład – porównywanie średnich

- Przy poziomie istotności α = 0.05 zweryfikuj hipotezę o równości średnich przy założeniu tej samej wariancji:

```
mean_1 = mean(x_1); mean_2 = mean(x_2)
s2_1 = var(x_1); s2_2 = var(x_2)
s2 = ((n_1 - 1)*s2_1 + (n_2 - 1)*s2_2)/(n_1 + n_2 - 2)
alfa = 0.05; c = qt(1 - alfa/2, n_1 + n_2 - 2)
T = abs(mean_1 - mean_2) / ( sqrt(s2*( n_1^(-1) + n_2^(-1) ) ) )
p_value = 2*(1 - pt(T, n_1 + n_2 - 2))
```

- Wykorzystanie funkcji pakietu R:

```
t.test(x_1, x_2, var.equal = T)
```

- Powtórz testy dla innych wartości mi_2, np. mi_2 = 0, 0.5, 2.

Przykład – porównywanie średnich

- Przeprowadź analogiczne testy bez zakładania równości wariancji

```
s2 = s2_1/n_1 + s2_2/n_2
d = (s2^2)/(((s2_1/n_1)^2)/(n_1-1) + ((s2_2/n_2)^2)/((n_2-1)))
c = qt(1 - alfa/2,  d)
T = abs(mean_1 - mean_2)/sqrt(s2)
p_value = 2*(1 - pt(T, d))
```

- Wykorzystanie funkcji pakietu R:

```
t.test(x_1, x_2)
```

Przykład – porównywanie średnich

- Przeprowadź analogiczne testy bez zakładania normalności rozkładów:

```
wilcox.test(x_1, x_2)
```

- Funkcja gęstości rozkładu Wilcoxona

```
plot(dwilcox(0:100, length(x_1), length(x_2)), type = 'l'); grid()
```

Przykład – test niezależności

- Przeprowadź testy niezależności dla danych zawartych w tabeli kontyngencji xx:

```
x_1 = c(16, 25, 11)
x_2 = c(13, 32, 15)
x_3 = c(31, 43, 26)
xx = cbind(x_1, x_2, x_3)
I = 3
J = 3
n_i = x_1 + x_2 + x_3
n_j = c(sum(x_1), sum(x_2), sum(x_3))
N = sum(n_j)
```

Przykład – test niezależności

- Obliczenie wartości statystyki decyzyjnej, progu i p-wartości:

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

- Wykorzystanie funkcji pakietu R:

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

Zadanie 1

- Plik tempciala.txt zawiera zarejestrowane wartości tętna oraz temperatury ciała dla 65 mężczyzn (płeć = 1) i 65 kobiet (płeć = 2).
- Dane zawarte w pliku można wczytać w następujący sposób:

```
dane = read.csv("tempciala.txt", header = T)
```

- Wyestymuj wartość średnią i odchylenie standardowe temperatury, osobno dla mężczyzn i kobiet.
- Osobno dla mężczyzn i kobiet przeprowadź:
  - testy tego, że średnia temperatura jest równa 36.6 ºC wobec hipotezy alternatywnej, że średnia temperatura jest inna, przyjmując, że rozkłady temperatur mają rozkład normalny,
  - testy normalności dla zarejestrowanych temperatur.

Zadanie 2

- W tabeli przedstawionej na następnym slajdzie zawarto dane dot. liczby samobójstw w Stanach Zjednoczonych w 1970 roku z podziałem na poszczególne miesiące.
- Zweryfikuj czy zamieszczone w niej dane wskazują na sezonową zmienność liczby samobójstw, czy raczej świadczą o stałej intensywności badanego zjawiska.

Zadanie 2 c.d.

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

Zadanie 3

- Dla wybranej spółki notowanej na GPW wczytaj dane ze strony bossa.pl.
  - Oblicz wartości procentowych zmian najniższych cen w poszczególnych dniach roku 2018. 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 2018 mają rozkład normalny.

Zadanie 4

- 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.
- *Oszacuj prawdopodobieństwo tego, że łożysko wykonane z pierwszego materiału będzie pracowało dłużej niż łożysko wykonane z materiału drugiego.

Zadanie 5

- Korzystając z danych zawartych na stronie pl.fcstats.com (zakładka *Porównanie lig*) zweryfikuj hipotezę o niezależności wyników (zwycięstw, remisów i porażek) gospodarzy od kraju, w którym prowadzone są rozgrywki piłkarskie.
- Testy przeprowadź na podstawie danych z lig:
  - polskiej – Ekstraklasa,
  - angielskiej – Premier League,
  - hiszpańskiej – Primera Division,
  - niemieckiej – Bundesliga.