Analiza Danych
Podstawy Statystyczne
(ADPS)

Laboratorium 2

Przykład – przedziały ufności

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

  n = 30

  x = **rnorm**(n, mean = 1, sd = 2)

- Na podstawie próby losowej wyestymuj wartość średnią oraz wariancję rozkładu generującego.

  x_mean = **mean**(x)

  x_var = **var**(x)

Przykład – przedziały ufności

- Zakładając, że wartość średnia i wariancja rozkładu generującego są nieznane wyznacz 90%, 95% i 99% przedziały ufności dla wartości średniej i wariancji.

  lev = 0.9 # 0.95, 0.99

  S = sd(x)

  w = S***qt**((1+lev)/2, n-1)/sqrt(n)

  ci_mean = c(x_mean - w, x_mean + w)

  a = (1 - lev)/2

  b = (1 - lev)/2

  ci_var = c((n-1)*S^2/**qchisq**(1-b,n-1), (n-1)*S^2/**qchisq**(a,n-1))

Przykład – bootstrap parametryczny

- Metodą bootstrapu **parametrycznego** oszacuj odchylenia standardowe estymatora wartości średniej i estymatora wariancji.

```
K = 1000
boot_res = replicate(K, {
    boot_dane = rnorm(n, mean = x_mean, sd = sqrt(x_var))
    c(mean(boot_dane), var(boot_dane))
    } )
sd_mean = sd(boot_res[1,])
sd_var = sd(boot_res[2,])
```

Przykład – bootstrap parametryczny

- Inny sposób – pętla for().

```
K = 1000
boot_mean = c()
boot_var = c()
for (k in 1:K) {
    boot_dane = rnorm(n, mean = x_mean, sd = sqrt(x_var))
    boot_mean[k] = mean(boot_dane)
    boot_var[k]  = var(boot_dane)
    }
sd_mean = sd(boot_mean)
sd_var = sd(boot_var)
```

Przykład – bootstrap nieparametryczny

- Metodą bootstrapu **nieparametrycznego** oszacuj odchylenia standardowe estymatora wartości średniej i estymatora wariancji.

```
K = 1000
boot_res = replicate(K, {
    boot_dane = sample(x, n, replace = T)
    c(mean(boot_dane), var(boot_dane))
    })
sd_mean = sd(boot_res[1,])
sd_var = sd(boot_res[2,])
```

Przykład – bootstrap nieparametryczny

- Inny sposób – pętla for().

```
K = 1000
boot_mean = c()
boot_var = c()
for (k in 1:K) {
    boot_dane = sample(x, n, replace = T)
    boot_mean[k] = mean(boot_dane)
    boot_var[k]  = var(boot_dane)
    }
sd_mean = sd(boot_mean)
sd_var = sd(boot_var)
```

Przykład – bootstrap

- Korzystając z wyników bootstrapu wyznacz przedziały ufności na poziomie 95%.

```
lev = 0.95
int_mean = quantile(boot_res[1,], c((1-lev)/2,(1+lev)/2))
int_var  = quantile(boot_res[2,], c((1-lev)/2,(1+lev)/2))
```

Przykład – metoda momentów

- Wygeneruj 1000 liczb będących próbą losową z rozkładu gamma o parametrach: shape = 1.1, scale = 5.

```
n = 1000
x_g = rgamma(n, shape = 1.1, scale = 5)
```

- Korzystając ze wzorów na estymatory parametrów rozkładu gamma dla metody momentów wyznacz ich wartości.

```
m1 = mean(x_g)
m2 = mean(x_g^2)
alpha_mom = m1^2/(m2 - m1^2)
beta_mom = (m2 - m1^2)/m1
```

Przykład – metoda najw. wiarygodn.

- Dla wygenerowanej próby losowej wyznacz wartości estymatorów parametrów uzyskane za pomocą metody największej wiarygodności. W celu uzyskania estymatora parametru kształtu α rozwiązujemy numerycznie równanie podane na wykładzie.

```
fun = function(x) digamma(x) - log(x) - mean(log(x_g)) +
   log(mean(x_g))

alpha_nw = uniroot(fun, lower = 0.5, upper = 4)$root

beta_nw = mean(x_g)/alpha_nw
```

Przykład – metoda najw. wiarygodn.

- Inny sposób z wykorzystaniem funkcji *fitdistr()* z pakietu MASS (należy dodać pakiet MASS).

```
est_nw = fitdistr(x_g, 'gamma', list(shape=1, scale=1), lower=0)
alpha_nw = as.numeric(est_nw$estimate[1])
beta_nw  = as.numeric(est_nw$estimate[2])
```

Zadanie 1

- Rozkład Poissona jest często używany do modelowania ruchu ulicznego (o małym natężeniu). Plik skrety.txt zawiera liczby pojazdów skręcających na pewnym skrzyżowaniu w prawo w przeciągu trzystu 3-minutowych przedziałów czasu (dane zostały zebrane o różnych porach dnia).
- Wczytaj dane za pomocą komendy *scan('skrety.txt').*
- Dopasuj do danych rozkład Poissona - wyestymuj parametr λ.
- Sprawdź zgodność otrzymanego rozkładu z zaobserwowanymi danymi porównując graficznie rzeczywiste (zaobserwowane) i spodziewane liczby „skrętów w prawo” (użyj funkcji *dpois()*).
- Metodą bootstrapu nieparametrycznego oszacuj odchylenie standardowe estymatora parametru λ.

Zadanie 2 – przedziały ufności

- Dla wybranej jednej spółki notowanej na GPW oblicz wartości procentowych zmian najwyższych cen w dniu i wykreśl ich histogram – zweryfikuj zgrubnie, czy możemy przyjąć, że procentowe zmiany cen otwarcia mają rozkład normalny.
- Wyestymuj wartość średnią oraz wariancję procentowych zmian najwyższych cen w dniu dla wybranej spółki.
- Zakładając, że zmiany cen otwarcia wartość mają rozkład normalny wyznacz 90%, 95% i 99% przedziały ufności dla wartości średniej i wariancji procentowych zmian najwyższych cen w dniu dla wybranej spółki.

Zadanie 3 – estymacja bayesowska

- Rzucona pinezka upada ostrzem do dołu lub do góry. Doświadczenie to można opisać rozkładem Bernoulliego z parametrem p będącym prawdopodobieństwem tego, że pinezka upadnie ostrzem do góry. Rozkład parametru p można opisać rozkładem Beta o parametrach α i β.
- Zaproponuj parametry rozkładu a priori parametru p oraz określ wartość oczekiwaną tego rozkładu.
- Rzuć pinezką 20 razy i zanotuj wyniki kolejnych rzutów. Wyznacz i narysuj rozkład a posteriori parametru p oraz oblicz wartość bayesowskiego estymatora $\hat{p}$.
  W rozważanym przypadku rozkład aposteriori parametru p jest również rozkładem Beta o parametrach:

$$\alpha_{post} = \alpha_{pr} + \sum_{i=1}^{n} x_i, \quad \beta_{post} = \beta_{pr} + n - \sum_{i=1}^{n} x_i, \quad x_i = \{0,1\}.$$

Zadanie 3 – estymacja bayesowska

- Rzuć pinezką jeszcze 20 razy i zanotuj wyniki.
  Wyznacz i narysuj rozkład a posteriori oparty na wszystkich 40 rzutach oraz oblicz wartość bayesowskiego estymatora $\hat{p}$.
  Porównaj wyniki z wynikami uzyskanymi po pierwszych 20 rzutach.
- Korzystając ze wzoru na wariancję rozkładu Beta wyznacz i porównaj wariancje rozkładu a priori, a posteriori po 20 rzutach i a posteriori po 40 rzutach.

Zadanie 4

- Plik fotony.txt zawiera odstępy między chwilami rejestracji kolejnych fotonów promieniowania gamma wykonywanymi za pomocą teleskopu kosmicznego Comptona (CGRO) w roku 1991.
- Metodą momentów oraz metodą największej wiarygodności wyznacz estymatory parametrów rozkładu gamma odpowiadające zarejestrowanym danym.
- Narysuj na jednym wykresie histogram odstępów oraz funkcje gęstości rozkładu gamma o parametrach wyestymowanych za pomocą obu metod.
- Metodą bootstrapu parametrycznego wyznacz odchylenia standardowe estymatorów parametrów oraz przedziały ufności na poziomie ufności 95%.

Dziękuję za uwagę!

Pytania ?