Rozkłady estymatorów

Rozkład chi-kwadrat

- **Definicja:**
  Niech $X_1, X_2, \dots, X_n$ będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie $N(0,1)$.
  Mówimy, że zmienna losowa
  $$X_1^2 + X_2^2 + \cdots + X_n^2$$
  ma rozkład **chi-kwadrat** z n stopniami swobody i oznaczamy $\chi^2(n)$.
- **Fakt:**
  Funkcja gęstości rozkładu $\chi^2(n)$ ma postać
  $$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} x^{n/2-1}\, e^{-x/2}, \qquad x > 0$$

Model wykładniczy

- Niech $X = (X_1, X_2, \dots, X_n)'$ będzie próbą z populacji o rozkładzie wykładniczym $Ex(\lambda)$ z parametrem $\lambda > 0$
- Estymator nieobciążony (o minimalnej wariancji) parametru λ ma postać

$$\hat{\lambda} = \frac{n-1}{n}\frac{1}{\bar{X}}$$

- **Fakt:**
  W modelu jednej próby prostej z rozkładu wykładniczego

$$2n\lambda\bar{X} \sim \chi^2(2n)$$

Model normalny

- Niech $X = (X_1, X_2, \dots, X_n)'$, $n > 1$ będzie próbą z populacji o rozkładzie normalnym $N(\mu, \sigma^2)$ z parametrami μ i σ^2
- Estymatory nieobciążone (o minimalnej wariancji) parametrów μ i σ^2 mają postać

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = S^2$$

- **Twierdzenie (Fishera)**
  W modelu jednej próby prostej z rozkładu normalnego

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

  Ponadto, estymatory $\bar{X}$ i S^2 są niezależnymi zmiennymi losowymi

Metoda Monte Carlo

- Niech $X = (X_1, X_2, \dots, X_n)'$ będzie próbą z populacji o rozkładzie P_θ z parametrem θ
- Ponadto, niech
$$\hat{\theta} = T(X)$$
będzie estymatorem parametru θ
- Załóżmy, że dysponujemy k niezależnymi realizacjami próby
$$X: x_1, x_2, .., x_n$$
oraz że
$$\hat{\theta}_i = T(x_i), \qquad i = 1,2, \dots, k$$
- **Fakt:**
Histogram wartości
$$\hat{\theta}_1 = T(x_1), \hat{\theta}_2 = T(x_2), \dots, \hat{\theta}_k = T(x_k)$$
jest dla dużych k dobrym przybliżeniem rozkładu $\hat{\theta}$

Metoda bootstrapowa

- Niech $X = (X_1, X_2, \ldots, X_n)'$ będzie próbą z populacji o rozkładzie P_θ z parametrem θ
- Ponadto, niech
$$\hat{\theta} = T(X)$$
będzie estymatorem parametru θ oraz F oznacza dystrybuantę rozkładu P_θ
- **Dystrybuantą empiryczną** nazywamy statystykę
$$\hat{F}(x) = \frac{\#\{k: X_k \leq x\}}{n}$$

Metoda bootstrapowa

- **Twierdzenie (Gliwenki-Cantelliego):**
  Niech $X = (X_1, X_2, \dots, X_n)'$ będzie próbą prostą z populacji o rozkładzie opisanym dystrybuantą F
  Wtedy zachodzi wzór

$$\sup_{-\infty < x < \infty} |\hat{F}(x) - F(x)| \xrightarrow{1} 0$$

Metoda bootstrapowa

- **Próbą bootstrapową** nazywamy próbę losową z rozkładu $\hat{F}$ i oznaczamy
$$X^* = (X_1^*, X_2^*, \ldots, X_n^*)'$$
- **Uwaga:**
W celu otrzymania realizacji próby bootstrapowej dokonujemy n-krotnego losowania **ze zwracaniem** spośród wartości oryginalnej próby
- **Fakt (zasada bootstrap):**
Rozkład statystyki
$$T(X^*) - \hat{\theta}$$
przy ustalonych wartościach $x_1, x_2, \ldots, x_k$ jest bliski rozkładowi
$$T(X) - \theta$$

Metoda bootstrapowa

- Załóżmy, że dysponujemy k realizacjami próby bootstrapowej

$$X^*: x_1^*, x_2^*, \dots, x_k^*$$

oraz że

$$\hat{\theta}_i^* = T(x_i^*), \qquad i = 1,2,\dots,k$$

- **Fakt:**
Histogram wartości

$$\hat{\theta}_i^* - \theta, \hat{\theta}_2^* - \theta, \dots, \hat{\theta}_k^*$$

jest dla dużych k dobrym przybliżeniem rozkładu

$$\hat{\theta} - \theta$$

Przedziały ufności

Przedziały ufności

- Niech $\theta \in \Theta$ oznacza parametr modelu statystycznego
- **Definicja:**
  Przedział (L, R) określony parą statystyk L i R takich, że
  $$P_\theta(L \leq R) = 1$$
  dla każdego $\theta \in \Theta$ nazywamy **przedziałem ufności** dla parametru θ na **poziomie ufności**
  $$1 - \alpha, \qquad 0 < \alpha < 1$$
  gdy dla każdego $\theta \in \Theta$ zachodzi nierówność
  $$P_\theta(L < \theta < R) \geq 1 - \alpha$$
- **Uwaga:** Typowe wartości poziomu ufności to: $0{,}9; 0{,}95; 0{,}99$ zazwyczaj podawane w procentach [%]

Konstrukcja przedziałów ufności

- **Definicja:**
  Funkcję $Q(X, \theta)$ nazywamy funkcją centralną dla parametru θ, gdy

  1. rozkład prawdopodobieństwa zmiennej losowej Q jest absolutnie ciągły i nie zależy od parametru θ

  2. funkcja $Q(X, \theta)$ jest ciągła i ściśle monotoniczna względem θ
- **Uwaga:**
  Warunek pierwszy można osłabić, żądając tylko by rozkład graniczny zmiennej losowej Q był absolutnie ciągły. Wtedy uzyskany przedział można stosować jedynie dla dużych prób

Konstrukcja przedziałów ufności

Konstrukcja przedziału ufności:

- Obieramy funkcję centralną $Q(X,\theta)$
- Wyznaczamy stałe a i b tak, aby
$$P(a < Q < b) = 1 - \alpha$$
- Rozwiązujemy nierówność
$$a < Q(X,\theta) < b$$
względem θ otrzymując szukany przedział
$$(L(X), R(X))$$
- **Uwaga:** Stałe a i b można dobrać na wiele sposobów. Zazwyczaj dobieramy tak, by
$$P(Q \le a) = P(Q \ge b) = \frac{\alpha}{2}$$

Przedziały ufności

- **Fakt:**
  W modelu jednej próby prostej z rozkładu wykładniczego, $(1-\alpha)\cdot 100\%$ przedział ufności dla parametru λ ma postać

$$\left(\frac{\chi^2\left(\frac{\alpha}{2}, 2n\right)}{2n\bar{X}}; \frac{\chi^2(1-\frac{\alpha}{2}, 2n}{2n\bar{X}}\right)$$

  gdzie $\chi^2(p,n)$ oznacza kwantyl rzędu p z rozkładu $\chi^2(n)$

Przedziały ufności

- **Definicja:**
  Niech $X \sim N(0,1)$ oraz $Y \sim \chi^2(n)$ będą niezależnymi zmiennymi losowymi.
  Mówimy, że zmienna losowa
  $$\frac{X}{\sqrt{\frac{1}{n}Y}}$$
  ma rozkład ***t*-Studenta** z n stopniami swobody i oznaczamy $t(n)$
- **Fakt:**
  Funkcja gęstości rozkładu t-Studenta ma postać
  $$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad x \in R$$

Przedziały ufności

- **Fakt:**
  W modelu jednej próby prostej z rozkładu normalnego $(1-\alpha)\cdot 100\%$ przedział ufności parametru μ ma postać

$$\left(\bar{X} - \frac{S}{\sqrt{n}}\, t\left(1-\frac{\alpha}{2}, n-1\right),\ \bar{X} + \frac{S}{\sqrt{n}}\, t\left(1-\frac{\alpha}{2}, n-1\right)\right)$$

gdzie $t(p,n) = F_t^{-1}(p)$ oznacza kwantyl rzędu p z rozkładu $t(n)$

Przedziały ufności

- **Fakt:**
  W modelu jednej próby prostej z rozkładu normalnego $(1-\alpha)\cdot 100\%$ przedział ufności parametru σ^2 ma postać

$$\left(\frac{(n-1)S^2}{\chi^2\left(1-\frac{\alpha}{2}, n-1\right)}; \frac{(n-1)S^2}{\chi^2\left(\frac{\alpha}{2}, n-1\right)}\right)$$

gdzie $\chi^2(p,n)$ oznacza kwantyl rzędu p z rozkładu $\chi^2(n)$

R

Funkcje związane z przedziałami ufności:

- **e...** (EnvStats) – pozwalają wyznaczyć przedziały ufności dla parametrów wybranego modelu
- Przykładowo:

  eexp (EnvStats) – pozwala wyznaczyć przedział ufności dla parametru w modelu wykładniczym

  enorm (EnvStats) – pozwala wyznaczyć przedziały ufności dla parametrów w modelu normalnym