# 6. WNIOSKOWANIE STATYSTYCZNE. PARAMETRYCZNA ESTYMACJA PUNKTOWA

We **wnioskowaniu statystycznym** z populacji pobieramy próbę i na jej podstawie wyciągamy wnioski dotyczące całej populacji. Bardzo ważny jest wybór owej próby, tak by zawierała jak najwięcej informacji o badanej populacji. Jedną z metod jest wybór tzw. *prostej próby losowej.*

**Definicja.** Jeśli $X_1, X_2, \ldots, X_n$ są niezależne i mają ten sam rozkład co cecha populacji $X$, to $X_1, X_2, \ldots, X_n$ nazywamy *(prostą) próbą losową* z $X$.

Oczywiście założenie, że pracujemy z prostą próbą losową, musi mieć swoje odzwierciedlenie podczas procesu zbierania danych - do próby powinniśmy wybierać niezależne od siebie obserwacje i każda z nich powinna dobrze reprezentować badaną populację.

W wyniku zebrania danych otrzymujemy *realizację próby losowej*, czyli $n$ ustalonych wartości, które oznaczamy $x_1, x_2, \ldots, x_n$.

Na podstawie próby losowej $X_1, X_2, \ldots, X_n$ chcemy opisać rozkład $X$. Możliwe są dwa podejścia.

1. **Podejście parametryczne** - zakładamy, że $X$ ma rozkład o dystrybuancie o znanej postaci a nie znamy jedynie parametrów tej dystrybuanty.
2. **Podejście nieparametryczne** - nie zakładamy, że $X$ ma rozkład, którego dystrybuanta należy do pewnej rodziny dystrybuant, indeksowanej skończenie wymiarowym parametrem.

W podejściu parametrycznym zakładamy zatem, że

$$\boxed{\begin{array}{l} X \sim F_\theta, \text{ gdzie } \theta \in \Theta \subseteq \mathbb{R}^k \\ \qquad\qquad \text{i} \\ X_1, X_2, \ldots, X_n \text{ jest próbą losową z } X \end{array}}$$

i szukamy nieznanego parametru $\theta$.

**Estymacja punktowa** polega na oszacowaniu $\theta$ za pomocą funkcji mierzalnej, której argumetami są elementy próby losowej $X_1, X_2, \ldots, X_n$. Oszacowanie takie będziemy nazywać estymatorem $\theta$ i oznaczać $\hat{\theta}$.

**Definicja.** *Estymatorem (punktowym)* $\theta$, oznaczanym $\hat{\theta}$, nazywamy dowolną funkcję mierzalną próby $t(X_1, X_2, \ldots, X_n)$, która służy do szacowania $\theta$.

## METODY WYZNACZANIA ESTYMATORÓW

Powstaje pytanie jak znajdować funkcję $t$ z definicji estymatora. Znanych jest wiele metod, poniżej przedstawiamy trzy z nich.

**1. Metoda momentów**

Jeśli $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ jest $k$-wymiarowym wektorem, to wyznaczamy $EX, EX^2, \ldots, EX^k$. Wszystkie te momenty zależą od nieznanych parametrów $\theta_1, \theta_2, \ldots, \theta_k$:

$$\begin{cases} EX &= \mu_1(\theta_1, \theta_2, \ldots, \theta_k) \\ EX^2 &= \mu_2(\theta_1, \theta_2, \ldots, \theta_k) \\ \vdots & \\ EX^k &= \mu_k(\theta_1, \theta_2, \ldots, \theta_k) \end{cases}.$$

Powyższy układ równań rozwiązujemy ze względu na $\theta_1, \theta_2, \ldots, \theta_k$ (zakładamy, że istnieje dokładnie jedno rozwiązanie):

$$\begin{cases} \theta_1 &= g_1(EX, EX^2, \ldots, EX^k) \\ \theta_2 &= g_2(EX, EX^2, \ldots, EX^k) \\ \vdots & \\ \theta_k &= g_k(EX, EX^2, \ldots, EX^k) \end{cases}.$$

Następnie momenty teoretyczne $EX, EX^2, \ldots, EX^k$ zamieniamy na momenty empiryczne $M_1$, $M_2$, …, $M_k$, gdzie $M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$. Uzyskane w ten sposób funkcje to szukane estymatory:

$$\begin{cases} \hat{\theta}_1 &= g_1(M_1, M_2, \ldots, M_k) \\ \hat{\theta}_2 &= g_2(M_1, M_2, \ldots, M_k) \\ \vdots & \\ \hat{\theta}_k &= g_k(M_1, M_2, \ldots, M_k) \end{cases}.$$

**Przykład 6.1.** Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu dwupunktowego z prawdopodobieństwem sukcesu $\theta$, gdzie $\theta \in (0, 1)$. Metodą momentów wyznaczymy estymator parametru $\theta$.

$$\hat{\theta} = \bar{X}.$$

**Przykład 6.2.** Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$. Wyznaczymy estymator parametru $(\mu, \sigma^2)$ stosując metodę momentów.

$$\hat{\mu} = \bar{X},\ \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Pojęcie *estymator wyznaczony metodą momentów* nie jest jednoznaczne. W wyżej opisanej metodzie momenty $EX, EX^2, \ldots, EX^k$ możemy zastąpić przez momenty centralne $E(X-EX)^2$, $E(X-EX)^3$, ..., $E(X-EX)^{k+1}$ a momenty empiryczne $M_1, M_2, \ldots, M_k$ - na centralne momenty empiryczne $\tilde{M}_2, \tilde{M}_3, \ldots, \tilde{M}_{k+1}$, gdzie $\tilde{M}_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$. Tak otrzymane estymatory też będziemy nazywać estymatorami wyznaczonymi metodą momentów. Ponadto zamiast kolejnych pierwszych momentów możemy użyć innych i znowu otrzymane estymatory nazwiemy estymatorami wyznaczonymi metodą momentów.

**2. Metoda kwantyli**

Jest to metoda analogiczna do metody momentów. Różnica polega na tym, że zamiast wyznaczać momenty $EX, EX^2, \ldots, EX^k$ i potem zastępować je momentami empirycznymi $M_1$, $M_2$, ..., $M_k$, wyznaczamy $k$ kwantyli różnych rzędów $X$ i potem zastępujemy je kwantylami empirycznymi (czyli kwantylami wyznaczonymi z próby).

Kwantyl empiryczny rzędu $p \in (0,1)$, oznaczany $q_p$, to wartość taka, że mniej-więcej $p \cdot 100\%$ obserwacji jest $\leqslant q_p$ i mniej-więcej $(1-p)\cdot 100\%$ obserwacji jest $\geqslant q_p$. My dla ustalenia uwagi kwantyl rzędu $p$ będziemy wyznaczać używając wzoru:

$$q_p = \begin{cases} \frac{x_{np:n} + x_{np+1:n}}{2} & \text{jeśli } np \text{ jest liczbą całkowitą} \\ x_{\lceil pn \rceil : n} & \text{jeśli } np \text{ nie jest liczbą całkowitą} \end{cases},$$

gdzie $x_{i:n}$ to $i$-ta obserwacja po ustawieniu wszystkich obserwacji w kolejności niemalejącej;

**3. Metoda największej wiarygodności**

Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu o gęstości $f_\theta(x)$, zaś $x_1, x_2, \ldots, x_n$ - jej realizacją. Wtedy funkcję

$$L(\theta; x_1, x_2, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n)$$

nazywamy **funkcją wiarygodności** rozważanego eksperymentu.

Analogicznie, jeśli $X_1, X_2, \ldots, X_n$ jest prostą próbą losową z rozkładu o masie prawdopodobieństwa $p_\theta(x)$, zaś $x_1, x_2, \ldots, x_n$ - jej realizacją, to funkcją wiarygodności nazywamy

$$L(\theta; x_1, x_2, \ldots, x_n) = p_\theta(x_1) p_\theta(x_2) \ldots p_\theta(x_n).$$

**Definicja.** *Estymator największej wiarygodności*, oznaczany $\hat{\theta}_{NW}$, to wartość parametru $\theta$, która, przy ustalonej realizacji próby $x_1, x_2, \ldots, x_n$, maksymalizuje funkcję wiarygodności:

$$\hat{\theta}_{NW} = \arg\max_{\theta \in \Theta} L(\theta; x_1, x_2, \ldots, x_n). \tag{1}$$

Dla uproszczenia rachunków, maksymalizację funkcji wiarygodności często warto zastąpić maksymalizacją jej logarytmu naturalnego:

$$\hat{\theta}_{NW} = \arg\max_{\theta\in\Theta} \ln L(\theta; x_1, x_2, \ldots, x_n). \tag{2}$$

Ponieważ funkcja $f(x) = \ln x$ jest ściśle rosnąca, wzory (1) i (2) są równoważne.

**Przykład 6.3.** Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu dwupunktowego z prawdopodobieństwem sukcesu $\theta$, gdzie $\theta \in (0, 1)$. Metodą największej wiarygodności wyznaczymy estymator parametru $\theta$.

$$\hat{\theta}_{NW} = \bar{X}.$$

**Przykład 6.4.** Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$. Wyznaczymy estymator największej wiarygodności parametru $(\mu, \sigma^2)$.

$$\hat{\mu}_{NW} = \bar{X},\ \hat{\sigma^2}_{NW} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

**Twierdzenie 6.1.** Jeśli $\hat{\theta}$ jest estymatorem największej wiarygodności parametru $\theta$ i $g$ jest funkcją mierzalną, to $g(\hat{\theta})$ jest estymatorem największej wiarygodności parametru $g(\theta)$.

**Przykład 6. 5.** Pokazaliśmy, że w przypadku próby losowej z rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$ mamy

$$\hat{\mu}_{NW} = \bar{X},\ \hat{\sigma^2}_{NW} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Z powyższego twierdzenia natychmiast wynika,że estymatorem największej wiarygodności parametru $(\mu, \sigma)$ jest

$$\hat{\mu}_{NW} = \bar{X},\ \hat{\sigma}_{NW} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Metoda największej wiarygodności jest zaimplementowana w R.

```
> install.packages("MASS")
> library("MASS")
> fitdistr(x, densfun, start)
```

gdzie

- `x` to realizacja próby losowej,
- `densfun` to nazwa rozkładu, np. densfun=”normal” , ”log-normal”, ”Poisson”, ”geometric”, ”exponential”,
- `start` to lista z początkowymi ocenami parametrów rozkładu, np.

  ```
  > fitdistr(x=dane, densfun="gamma", start=list(shape=3,rate=3));
  ```

  listy tej nie należy podawać dla rozkładów: ”normal” , ”log-normal”, ”Poisson”, ”geometric”, ”exponential” i innych, dla których znane są dokładne wzory na estymatory parametrów otrzymane metodą największej wiarygodności; dla pozostałych rozkładów stosuje się metody aproksymacyjne wyznaczania tych estymatorów i wtedy listę z początkowymi ocenami parametrów można podawać,
- `lower` i `upper` to dolne i górne ograniczenia na parametry rozkładu, warto je podawać, gdy używana jest aproksymacyjna metoda wyznaczania estymatorów tych parametrów; np. w przypadku rozkładu gamma zarówno parametr kształtu jak i drugi parametr muszą być dodatnie, zatem argument `lower` to wektor złożony z dwóch zer:

  ```
  lower=c(0,0)
  ```

## WIELOŚĆ ESTYMATORóW

Szacując nieznany parametr $\theta$ na podstawie próby losowej $X_1, X_2, \ldots, X_n$ z populacji $X \sim F_\theta$, $\theta \in \Theta$, możemy utworzyć wiele estymatorów.

**Przykład 6. 6.** Niech $X_1, X_2, \ldots, X_n$ będzie próbą losową z populacji $X$ o rozkładzie jednostajnym na przedziale $[0, \theta]$, gdzie $\theta > 0$, tzn. rozkładem o gęstości

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{dla } x \in [0, \theta] \\ 0 & \text{dla } x \notin [0, \theta] \end{cases}.$$

Szukamy estymatora parametru $\theta$.

I-szy sposób: metoda momentów

Otrzymujemy następujący estymator

$$\hat{\theta}_{MM} = 2\bar{X},$$

gdzie indeks $MM$ oznacza, że estymator został wyznaczony metodą momentów.

II-gi sposób: metoda największej wiarygodności

Otrzymujemy

$$\hat{\theta}_{NW} = \max(X_1, X_2, \ldots, X_n) =: X_{n:n}.$$

W powyższym przykładzie estymator otrzymany metodą momentów i estymator największej wiarygodności nie są takie same. Sytuacje, gdy do wyboru mamy różne estymatory, nie są odosobnione. Potrzebujemy zatem narzędzi pozwalających ocenić, który z konkurencyjnych estymatorów jest lepszy. Zostaną one omówione w następnym wykładzie.