Wykład 2: 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**. Aby zdefiniować to pojęcie przyjrzyjmy się bliżej postawionemu problemowi.

Niech X oznacza badaną cechę populacji. Na przykład

- populacją może być zbiór wszystkich 10-cio letnich dzieci mieszkających w Polsce, a X – wzrostem dziecka;
- populacją mogą być wszystkie żarówki energooszczędne produkowane przez pewnien zakład, a X - czasem świecenia żarówki;
- populacją mogą być wszystkie szklane abażury produkowane przez pewnien zakład, a X – informacją czy abażur posiada wady czy nie.

X jest zmienną losową, bo jego wartość zależy od zdarzenia losowego: w przykładzie ze wzrostem 10-cio letnich dzieci X zależy od wybranego dziecka; w przykładzie z żarówkami X zależy od wybranej żarówki. Naszym celem jest opisanie rozkładu X. Aby go osiągnąć, pobieramy próbę, którą oznaczamy

$$X_1, X_2, \ldots, X_n.$$

Przed zebraniem danych elementy próby to zmienne losowe. Zakładamy o nich, że mają ten sam rozkład, co badana cecha populacji X. Jeśli dodatkowo przyjmiemy, że są one niezależne, to będziemy mieć prostą próbę losową, często zwaną po prostu próbą losową.

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. Na przykład zakładamy, że
   - X ma rozkład wykładniczy z parametrem λ, $Exp(\lambda)$, gdzie $\lambda > 0$, i nie znamy jedynie wartości parametru λ;
   - X ma rozkład normalny o średniej $\mu \in \mathbb{R}$ i wariancji $\sigma^2 > 0$, $\mathcal{N}(\mu, \sigma^2)$ i nie znamy jedynie parametrów μ i σ^2.
2. **Podejście nieparametryczne** – nie zakładamy, że X ma rozkład, którego dystrybuanta należy do pewnej rodziny dystrybuant, indeksowanej skończenie wymiarowym parametrem.

Najpierw skupimy się na podejściu parametrycznym. Będziemy zatem zakładać, że X ma rozkład o dystrybuancie F z nieznanym parametrem θ, co symbolicznie zapisujemy $X \sim F_\theta$, gdzie θ może być zarówno jedno- jak i wielowymiarowym parametrem. Ponadto Θ oznaczać będzie zbiór wszystkich możliwych wartości parametru θ: $\theta \in \Theta$.

Reasumując, nasze założenia to:

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

Przy tych założenich przyjrzymy się dokładniej dwóm aspektom wnioskowania statystycznego:

- estymacji punktowej,
- weryfikacji hipotez.

Parametryczna estymacja punktowa

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

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

Powstaje pytanie jak znajdować funkcę t z definicji estymatora. Znanych jest wiele metod, poniżej przedstawiamy jedną z nich, uznawaną za najlepszą. Jest to tzw. **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 θ, 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 2. 1. Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu dwupunktowego z prawdopodobieństwem sukcesu θ, gdzie $\theta \in (0, 1)$. Metodą największej wiarygodności wyznaczymy estymator parametru θ.

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

Przykład 2.2. 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 (μ, σ^2).

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

Twierdzenie 2.1. Jeśli $\hat{\theta}_{NW}$ jest estymatorem największej wiarygodności parametru θ i g jest funkcją mierzalną, to $g(\hat{\theta}_{NW})$ jest estymatorem największej wiarygodności parametru $g(\theta)$.

Przykład 2.3. 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 (μ, σ) 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, lower, upper)
```

gdzie

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

  ```
  > fitdistr(x=dane, densfun="gamma", start=list(shape=2,rate=0.7))
  ```

  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)
  ```