Elementy statystyki

STA - Wykład 2

dr hab. Waldemar Wołyński
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza

Model statystyczny

Model: jedna próba

Jeżeli próbka $\boldsymbol{x} = (x_1, \ldots, x_n)'$ jest **reprezentatywna**, to stanowi ona podstawę do wnioskowania o **populacji** z której pochodzi. Wnioskowanie takie wymaga zbudowania modelu "zachowania się" zmiennej X w populacji. Budowa modelu polega na przyjęciu założenia o rozkładzie (teoretycznym) zmiennej X w populacji oraz traktowaniu obserwacji jako wartości tej zmiennej.

Dokładniej: budując model statystyczny traktujemy wektor obserwacji (próbkę) $\boldsymbol{x} = (x_1, \ldots, x_n)'$ jak realizację wektora losowego (próby) $\boldsymbol{X} = (X_1, \ldots, X_n)'$ z nieznanego (lub jedynie częściowo nieznanego) rozkładu.

Statystyka, to każda (mierzalna) funkcja próby. Zatem w modelu, statystyka jest wielkością losową.

Model: jedna próba prosta

Modelujemy wyniki doświadczenia w którym dokonujemy n-**niezależnych** obserwacji badanej cechy X na losowo wybranych z populacji (jednorodnej ze względu na badana cechę) jednostkach eksperymentalnych.

Przykład 1. W celu określenia czasu bezawaryjnej pracy urządzeń po wykonaniu kapitalnego remontu, wybrano 50 urządzeń i obserwowano czas ich bezawaryjnej pracy. Wyniki (w h.) są następujące: 629, 325, 215, ... ,612, 841, 492. Budując model statystyczny tego eksperymentu zakładamy, że (w populacji) czas bezawaryjnej pracy urządzenia (cecha X) ma rozkład wykładniczy z nieznanym parametrem λ.

Model ten ma jeden parametr: λ.

Przykład 2. Przeprowadzono 50 niezależnych eksperymentów polegających na hamowaniu badanego typu samochodu wyposażonego w nowy typ układu hamulcowego (na suchym asfalcie, przy prędkości 40 km/h, itd.). Notowano długość drogi hamowania z dokładnością do jednego centymetra. Otrzymane wyniki to: 18.66, 17.81, 18.96, ... ,17.62, 18.61, 17.99.
Budując model statystyczny tego eksperymentu zakładamy, że (w populacji) długość drogi hamowania (cecha X) ma rozkład normalny z nieznanymi parametrami μ i σ^2

Uwaga! Model ten często zapisujemy w następującej postaci:

$$X_i = \mu + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

X_i – i-ta obserwacja badanej cechy X,
μ – wartość oczekiwana (średnia, "prawdziwa" wartość) badanej cechy X,
ε_i – błędy (niezależne zmienne losowe o jednakowym rozkładzie $N(0, \sigma^2)$).

Model ten ma dwa parametry: μ i σ^2.

Estymacja punktowa

Metoda największej wiarogodności

Niech $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ będzie próbą z populacji o rozkładzie P_θ, gdzie $\theta \in \Theta \subset \mathbf{R}^d$.
Ponadto, niech rozkłady P_θ opisane będą za pomocą funkcji prawdopodobieństwa (gęstości) p_θ.

Definicja

Funkcję L określoną wzorem

$$L(\theta; \boldsymbol{x}) = p_\theta(\boldsymbol{x})$$

*nazywamy **funkcją wiarogodności**.*

Uwaga! Funkcją wiarogodności nazywamy czasem funkcję $\ln p_\theta(\boldsymbol{x})$.

Metoda największej wiarogodności

Definicja

Estymatorem największej wiarogodności *(ENW) parametru θ nazywamy statystykę $\hat{\theta}(\boldsymbol{X})$, której wartości $\hat{\theta}(\boldsymbol{x})$ spełniają warunek:*

$$\forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} : \; L(\hat{\theta}(\boldsymbol{x}); \boldsymbol{x}) = \sup_{\theta \in \Theta} L(\theta; \boldsymbol{x}).$$

Uwagi! 1) Dla danego parametru θ, ENW może nie istnieć lub może być wyznaczony niejednoznacznie.
2) Zazwyczaj, podczas wyznaczanie ENW, wygodniej jest operować funkcją $\ln L$ niż funkcją L.

Estymacja parametru λ w modelu wykładniczym

Do estymacji parametru λ (przykład 1) możemy użyć metody **największej wiarogodności**.

Fakt

Estymatorem największej wiarogodności (ENW) parametru λ, w modelu jednej próby prostej z rozkładu wykładniczego, jest statystyka

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

Estymacja parametru μ w modelu normalnym

Do estymacji parametru μ (przykład 2) możemy użyć metody **najmniejszych kwadratów**, polegającej na minimalizacji sumy kwadratów błędów, tzn.

$$S(\mu) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x_i - \mu)^2.$$

Fakt

Estymatorem najmniejszych kwadratów (ENK) parametru μ, w modelu jednej próby prostej z rozkładu normalnego, jest statystyka

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

Uwaga! Zwróćmy uwagę, że normalność rozkładu obserwacji (błędów) nie jest konieczna. Wystarczy jedynie założenie niezależności (a nawet nieskorelowania) błędów.

Estymatory nieobciążone

Estymatory nieobciążone

Niech $\theta \in \Theta$ oznacza parametr modelu statystycznego.

Definicja

Statystykę $\hat{\theta}$ nazywamy **estymatorem nieobciążonym** *parametru θ, gdy dla każdego $\theta \in \Theta$:*

$$E(\hat{\theta}) = \theta.$$

Uwaga! Klasa estymatorów nieobciążonych danego parametru może być pusta. Zazwyczaj jednak, dla danego parametru istnieje wiele różnych estymatorów nieobciążonych. Najlepszym z nich jest ten, który ma minimalną wariancję - estymator nieobciążony o minimalnej wariancji (ENMW).

Estymatory nieobciążone

Fakt

W modelu jednej próby prostej z rozkładu wykładniczego, ENW parametru λ postaci

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

jest obciążonym estymatorem tego parametru.
Estymator nieobciążony (o minimalnej wariancji) parametru λ ma postać:

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

Estymatory nieobciążone

Fakt

W modelu jednej próby prostej z rozkładu normalnego, ENK parametru μ postaci

$$\hat{\mu} = \bar{X}$$

jest nieobciążonym (o minimalnej wariancji) estymatorem tego parametru.
Ponadto, statystyka

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

jest nieobciążonym (o minimalnej wariancji) estymatorem parametru σ^2.

Rachunek prawdopodobieństwa

Definicja

Niech $X_1, X_2, \ldots, X_n$ 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 (ozn. $\chi^2(n)$).*

Fakt

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-(x/2)}, \quad x > 0.$$

Twierdzenie Fishera

Twierdzenie

W modelu jednej próby prostej z rozkładu normalnego estymatory

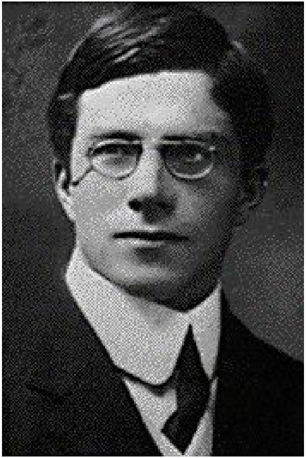
$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \quad i \quad S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

są niezależnymi zmiennymi losowymi oraz

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad i \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Trochę historii ...

Sir Ronald A. Fisher
(1890 - 1962)



Fisher w wieku 22 lat