

3.1 Wprowadzenie do estymacji

Ile mamy czerwonych krwinek w krwi? Ile karpia żyje w odrze? Ile ton trzody chlewnej będzie wyprodukowane w przyszłym roku? Ile białych samochodów jeździ ulicami Warszawy? Ile liści rośnie na najwyższym drzewie we Wrocławiu? W większości przypadków są to ciekawe pytania, w większości przypadków nie jesteśmy w stanie poznać prawdziwej odpowiedzi na to pytanie, w każdym przypadku możemy tą wartość ocenić. O sposobach oceny interesujących nas parametrów będzie poniżej.

Zacznijmy jednak od słownikowych definicji

Estymacja, to proces, którego celem jest ocena nieznanego wartości (funkcji) parametru na podstawie obserwacji.

Estymator, to funkcja służąca do oceny nieznanego wartości (funkcji) parametru.

Wartość estymatora, to ocena wartości (funkcji) parametru dla danej obserwacji.

Poniżej omówimy kilka sposobów konstrukcji estymatorów. W dalszej części znajdują się przykłady użycia, oraz zadania do wykonania. Więcej informacji o estymatorach pojawi się na zajęciach poświęconych regresji.

Rozmawiając z prowadzącym, dobrze rozróżniać estymator od wartości estymatora.

3.2 Konstrukcja estymatorów

Teoria estymacji to dział statystyki zajmujący się dwoma zagadnieniami: konstrukcją estymatorów oraz wykazywaniem ich właściwości. Nie jest to najważniejsze miejsce aby opisywać metody konstrukcji estymatorów (będzie na wykładzie, jest w polecanych książkach, pokażemy prosty przykład dla rozkładu gamma), osoby zainteresowane powinny szukać informacji o metodach (poniżej najpopularniejsze)

- metoda momentów,

Wartość oczekiwana (pierwszy moment) dla zmiennej o rozkładzie $\gamma(\alpha, \lambda)$ to

$$E(X) = \alpha\lambda,$$

drugi moment (wariancja) to

$$Var(X) = \alpha\lambda^2.$$

Powyższe wzory można przekształcić i sprowadzić do następującej postaci

$$\begin{aligned}\lambda &= \frac{Var(X)}{E(X)}, \\ \alpha &= \frac{(E(X))^2}{Var(X)}.\end{aligned}$$

Teraz możemy posłużyć się znanymi estymatorami średniej i wariancji, by wyznaczyć estymatory interesujących parametrów

$$\begin{aligned}\hat{\lambda} &= \frac{\widehat{Var}(X)}{\widehat{E}(X)}, \\ \hat{\alpha} &= \frac{(\widehat{E}(X))^2}{\widehat{Var}(X)}.\end{aligned}$$

Estymatory otrzymane tą metodą nie zawsze są dobre, dla rozkładu gamma stosowanie estymatorów wyznaczonych metodą momentów jest niepolecane (zadanie: zbadaj wariancję i obciążenie takiego estymatora). Zaletą takich estymatorów jest łatwość ich wyznaczenia. W znakomitej liczbie przypadków wystarczy wyznaczyć dwa pierwsze momenty. Bardziej zaawansowaną metodą wyznaczania estymatorów metodą momentów jest numeryczna aproksymacja parametrów rozkładu na bazie czterech pierwszych momentów z próby.

- metoda największej wiarygodności,
- metoda najmniejszych kwadratów,
- estymacja Bayesowska,
- metoda bootstrapowa.

3.3 Właściwości estymatorów

Nie jest sztuką powiedzieć, że za tydzień w poniedziałek będzie słonecznie. Sztuką jest mieć rację.

Każdy może zaproponować jakiś estymator, poniżej przedstawiamy zestaw własności, które powinien mieć dobry estymator. To, który estymator ma które własności będziemy a zajęciach. Wymienione własności to nie wszystkie możliwe własności, a tylko te najczęściej opisywane.

- Nieobciążoność. Estymator $T(X)$ jest nieobciążonym estymatorem funkcji $g(\theta)$ jeżeli

$$E_{\theta}[T(X)] = g(\theta),$$

czyli jeżeli wartość oczekiwana wartości estymatora jest równa wartości ocenianego parametru.

- Minimalna wariancja. Estymator $T(X)$ jest estymatorem o minimalnej wariancji w danej klasie estymatorów, jeżeli dla każdego θ ma najmniejszą wariancję, spośród estymatorów w danej klasie.

- **Dopuszczalność.** Estymator jest dopuszczalny w danej klasie estymatorów, jeżeli w tej klasie nie ma estymatora lepszego (w sensie błędu średniokwadratowego).
- **Normalność.** Estymator jest normalny, jeżeli rozkład wartości estymatora jest rozkładem normalnym. Ta własność przydaje się przy konstrukcji przedziałów ufności.
- **Zgodność.** Estymator jest zgodny, jeżeli z $n \rightarrow \infty$ estymator zbiega do prawdziwej wartości ocenianego parametru (wariancja i obciążenie estymatora zbiega do zera).

proponuje jeszcze „niezależność” - jeżeli wynik estymacji nie zależy od badacza.

3.4 Popularne estymatory

Przez $x = (x_1, \dots, x_n)$ oznaczmy wektor obserwacji. Poniżej przedstawiamy najpopularniejsze estymatory.

Średnia arytmetyczna

Średnia jest też pierwszym momentem z próby.

$$\bar{x} = \frac{1}{n} \sum (x_i),$$

Wariancja

Estymator wariancji gdy średnia jest znana

$$S_1^2 = \frac{1}{n} \sum (\bar{x} - x_i)^2,$$

Estymator wariancji gdy średnia jest nie znana (nieobciążony)

$$S_2^2 = \frac{1}{n-1} \sum (\bar{x} - x_i)^2.$$

Odchylenie standardowe

$$\hat{\sigma} = \text{sqrt}(S^2)$$

Odchylenie średnie

$$d = \frac{1}{n} \sum |\bar{x} - x_i|$$

Współczynnik zmienności

$$V_s = \frac{\hat{\sigma}}{\bar{x}}$$

$$V_d = \frac{d}{\bar{x}}$$

Współczynnik skośności

$$W_S = \frac{\bar{x} - d}{\hat{\sigma}}$$

Współczynnik asymetrii

$$A = \frac{1}{n\hat{\sigma}^3} \sum (x_i - \bar{x})^3$$

Współczynnik kurtozy

$$A = \frac{1}{n\hat{\sigma}^4} \sum (x_i - \bar{x})^4$$

3.4.1 Miary pozycyjne**Dominana**

Najczęstsza wartość w próbie

Mediana

Wartość środkowego elementu (jeżeli elementów jest nieparzysta liczba), lub średniej z dwóch elementów najbliższych środka (jeżeli elementów jest parzysta liczba).

Kwantyl

Kwantyl rzędu p to wartość $p * n$ tej statystyki pozycyjnej z próby.

Percentyle - kwantyle o rzędach będących wielokrotnością 0.01.

Kwartyle - kwantyle o rzędach będących wielokrotnością 0.25 (są trzy, górny oznaczany Q_3 , dolny oznaczany Q_1 i środkowy).

Rozstęp

$$R = \max(x) - \min(x)$$

Rozstęp kwartyłowy

$$Q_{1,3} = Q_3(x) - Q_1(x)$$

3.4.2 Miary zależności**Kowariancja**

$$Cov(x, y) = \frac{1}{n-1} \sum_i \sum_j (x_i y_j - \bar{x} \bar{y})$$

Korelacja Pearsona

$$Cor(x, y) = \frac{cov(x, y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

Korelacja rang Spearmana

$$r_s = 1 - \frac{6 \sum_i r_i}{n(n^2 - 1)}$$

gdzie r_i - różnica pomiędzy rangą elementu x_i i y_i .

Korelacja rang Kendalla

$$\tau = \frac{2(N^+ - N^-)}{n(n-1)}$$

gdzie N^+ liczba zgodnych par, czyli takich par (i, j) że (x_i, x_j) są w tej samej relacji (większe lub mniejsze) co (y_i, y_j) . N^- to liczba par niezgodnych.

3.5 Zadania:

1. Czy średnia z próby jest nieobciążonym estymatorem parametru średniej (położenia) dla rozkładów
 - normalnym,
 - log-normalnym,
 - cauchego.
2. Czy mediana z próby jest nieobciążonym estymatorem mediany dla rozkładów
 - normalnym,
 - log-normalnym,
 - cauchego.
3. Czy estymator wariancji $S_1 = \frac{1}{n} \sum (x - \bar{x})^2$ jest estymatorem nieobciążonym dla rozkładów
 - normalnym,
 - log-normalnym,
 - cauchego.

Porównać z estymatorem $S_2 = \frac{1}{n-1} \sum (x - \bar{x})^2$.

4. Wyznacz metodą momentów estymatory parametrów rozkładu gamma, następnie zbadaj czy są to estymatory nieobciążone.
5. Wyznacz 95% przedział ufności dla estymatora średniej dla rozkładu normalnego dla $n=20$ obserwacji.
6. Wyznacz 95% przedział ufności dla parametru p w rozkładzie dwumianowym.
7. Niech $f : R \rightarrow R$ będzie monotoniczną i odwracalną funkcją, a $T(X)$ będzie nieobciążonym estymatorem parametru p . Czy $f(T(X))$ jest nieobciążonym estymatorem parametru $f(p)$? Czy i kiedy taka zależność zachodzi? Czy zachodzi dla $f(x) = \sqrt{x}$?
8. Wyniki jednego z ostatnich sondaży przeprowadzonego przed wyborami prezydenckimi były następujące: 52% poparcia dla Donalda Tuska i 48% poparcia dla Lecha Kaczyńskiego. Wiedząc, że w sondażu uczestniczyło 1234 osób, wyznacz 95% przedział ufności dla parametru „poparcie dla Donalda Tuska”. Zinterpretuj wyniki.

9. * Obserwujesz kolejne realizacje zmiennej losowej o rozkładzie normalnym o nieznannej wartości średniej i wariancji. Ile obserwacji musisz zarejestrować, aby 95% przedział ufności dla oceny wariancji był węższy niż 0.1?
10. * Czy można symulacyjnie wykazać asymptotyczne nieobciążenie?
11. * Napisz program do wyznaczania bootstrapowego estymatora średniej. Sprawdź czy ten estymator jest nieobciążony.
12. * Jak mała może być wariancja estymatora? Jak mała może być wariancja estymatora nieobciążonego?