WYKŁAD 1 – cz.II

Ewa Frątczak

Wprowadzenie do przedmiotu.

Rozkłady statystyk z próby. Metody estymacji, własności estymatorów.

Struktura:

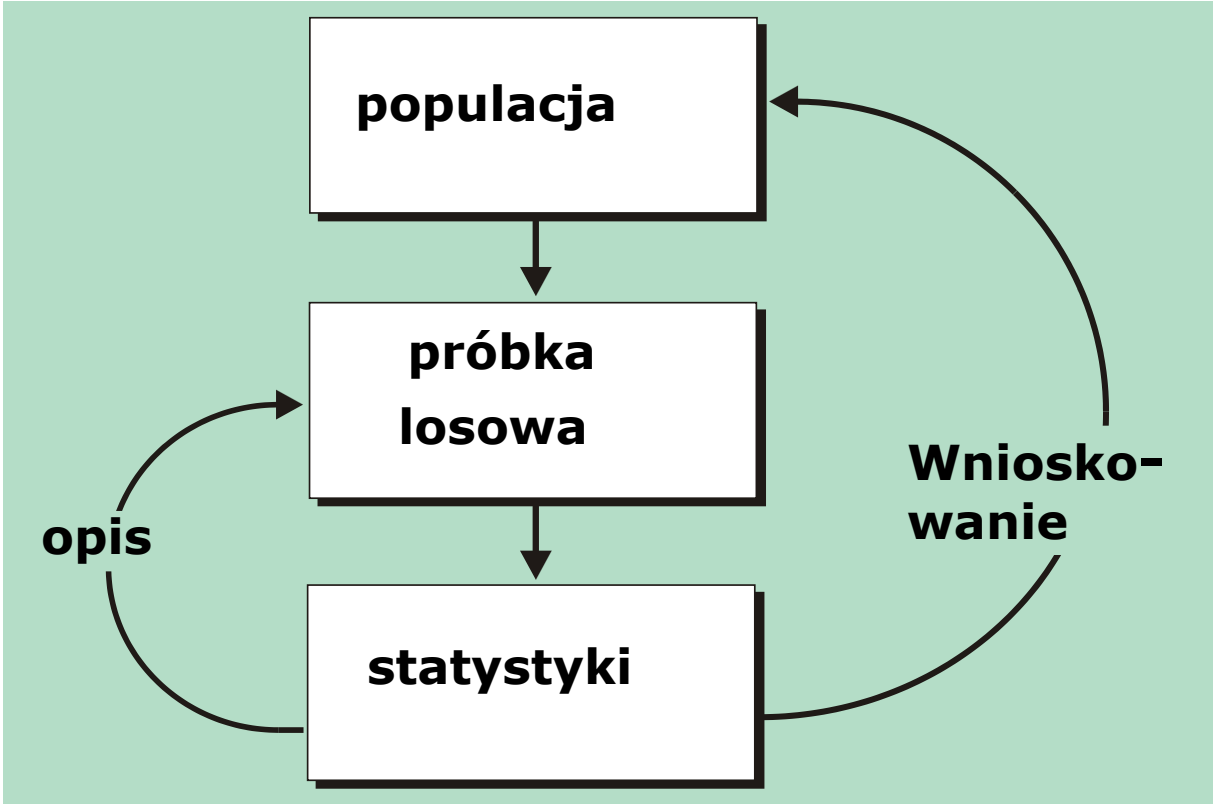
1. Harmonogram zajęć, literatura, zespół realizujący zajęcia , warunki zaliczenia przedmiotu
2. Istota wnioskowania statystycznego – podstawowe pojęcia
3. Statystyki z próby i ich rozkłady
3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady
3.2. Rozkłady dokładne statystyk z próby
3.3. Rozkłady graniczne statystyk z próby
4. Estymatory – własności – metody pozyskiwania
4.1. Podstawowe pojęcia
4.2. Podstawowe własności estymatorów
4.3. Metody pozyskiwania estymatorów
4.4. Przykład wyznaczania estymatora metodą MNW

1. Harmonogram zajęć, literatura, zespół realizujący zajęcia, warunki zaliczenia przedmiotu

Prowadzący zajęcia część II:

dr hab. prof. SGH, Ewa Frątczak
dr Wioletta Grzenda
dr Aneta Ptak-Chmielewska

2. Istota wnioskowania statystycznego – podstawowe pojęcia



2. Istota wnioskowania statystycznego – podstawowe pojęcia

Podstawowe pojęcia:

zbiorowość generalna (populacja), próba losowa, narzędzia wnioskowania

Podstawowe znaczenie dla teorii wnioskowania statystycznego (estymacji i weryfikacji hipotez statystycznych) posiada koncepcja **próby losowej** i **statystyki z próby**.

2. Istota wnioskowania statystycznego – podstawowe pojęcia

Wyróżnia się dwa podstawowe schematy losowania:

•**losowanie bez zwracania**: raz wylosowany element nie bierze udziału w dalszym losowaniu. Jest to losowanie zależne, a otrzymana w ten sposób próba jest próbą bez powtórzeń

•**losowanie ze zwracaniem**: raz wylosowany element wraca do populacji i może ponownie brać udział w losowaniu. Jest to losowanie niezależne, w wyniku którego otrzymuje się próbę z powtórzeniami.

Z reguły wnioskowanie statystyczne wymaga, aby próba była pobrana metodą ze zwracaniem tzn. wylosowana w sposób niezależny.

2. Istota wnioskowania statystycznego – podstawowe pojęcia

Jeśli rozpatruje się zmienną losową, która w populacji generalnej ma rozkład dany dystrybuantą F(x), to **n-elementową próbę losową** zapiszemy jako ciąg n-niezależnych zmiennych losowych $(X_1, X_2,...,X_n)$, z których każda ma taki sam rozkład dany dystrybuantą F(x).

Jeśli dokonuje się n-krotnych obserwacji, to otrzymamy konkretną realizację próby losowej, którą można zapisać jako: $(x_1, x_2,...,x_n)$ gdzie $x_1, x_2,...,x_n$ są wartościami odpowiadającymi poszczególnym elementom próby. W obu wypadkach używa się pojęcia próba.

Zbiór wszystkich możliwych prób losowych, określa się mianem **przestrzeni próby losowej**.

Próba losowa (wyniki próby losowej) jest podstawą do wnioskowania o parametrach bądź rozkładzie zmiennej losowej w populacji generalnej.

3. Statystyki z próby i ich rozkłady
3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Narzędziem tego wnioskowania są funkcje zmiennych losowych $X_1, X_2, \ldots, X_n$, charakteryzujące próbę losową i nazywane **statystykami z próby**.

Ogólnie zapisywać je będziemy jako:

$$Z_n = f(X_1, X_2, \ldots, X_n)$$

Statystyką z próby jest np. średnia z próby , wariancja z próby definiowana jako s_n^2 lub S_n^2 , gdzie:

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$s_n^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

$$S_n^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$$

3. Statystyki z próby i ich rozkłady
3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Wyróżnia się dwa typy rozkładów: **dokładne i graniczne**.

DOKŁADNE ROZKŁADY STATYSTYK Z PRÓBY - dla ustalonego n należy określić rozkład statystyki $T_n = T(X_1, X_2, ..., X_n)$. *Z* reguły liczba obserwacji jest mała i mówi się o tzw. rozkładach małych prób.

GRANICZNE ROZKŁADY STATYSTYK Z PRÓBY – rozkład T_n , gdy $n \to \infty$ (rozkłady asymptotyczne)

3. Statystyki z próby i ich rozkłady
3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Zmienna losowa ciągła *X* ma **rozkład normalny**, jeśli jej funkcja gęstości określona jest wzorem postaci :

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \quad , -\infty < x < +\infty \quad ,$$

gdzie m i δ są parametrami rozkładu.

Rozkład normalny ze średnią $m=0$ i odchyleniem standardowym $\sigma=1$ nazywamy standardowym rozkładem normalnym i oznaczamy $N(0,1)$.

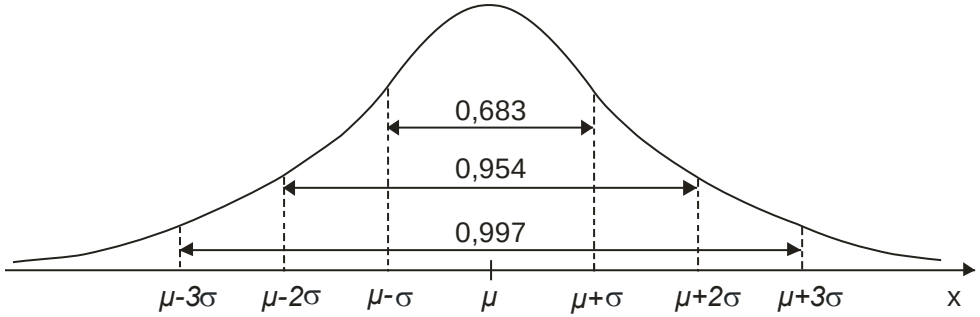
.

3. Statystyki z próby i ich rozkłady

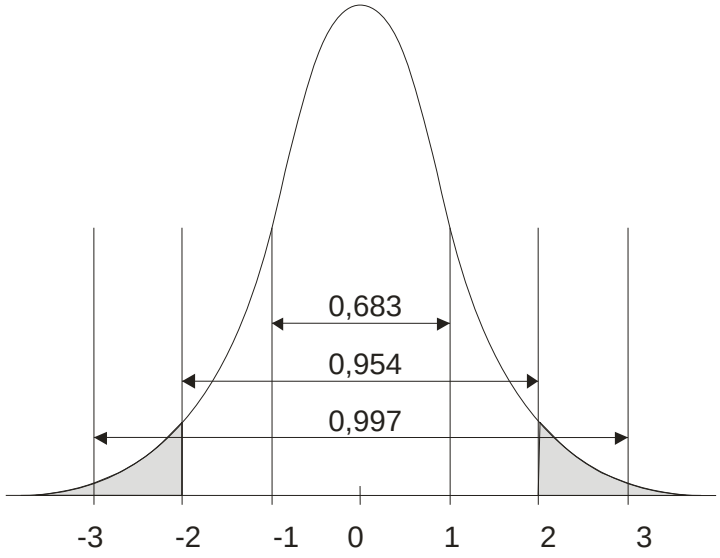
3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Wykresy funkcji gęstości rozkładu normalnego i normalnego standaryzowanego oraz reguła trzech sigm

$N(m,\sigma)$



$N(0,1)$



TWIERDZENIA O ROZKŁADZIE SUMY NIEZALEŻNYCH ZMIENNYCH LOSOWYCH

Twierdzenie 1.

Jeżeli zmienne losowe $X_1, X_2, ..., X_n$ są niezależne i zmienne losowa X_i dla $i = 1,2,...,n$ ma rozkład $N(m_i, \sigma_i)$ to zmienna losowa $Y = X_1 + X_2 + ... + X_n$ ma rozkład:

$$N\left(m_i + m_2 + ... + m_n, \sqrt{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2}\right)$$

Jeżeli zmienne losowe $X_1, X_2, ..., X_n$ są niezależne o takim samym rozkładzie ($X_i - N(m, \sigma)$ dla $i = 1,2,...,n$) to zmienna losowa *Y* ma rozkład $N\left(nm, \sigma\sqrt{n}\right)$.

TWIERDZENIA O ROZKŁADZIE SUMY NIEZALEŻNYCH ZMIENNYCH LOSOWYCH

Twierdzenie 2.

Jeżeli zmienne losowe $X_1, X_2, ..., X_n$ są niezależne i zmienne losowa X_i dla $i = 1,2,...,n$ ma rozkład $N(m_i, \sigma_i)$ to zmienna losowa $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ma rozkład:

$$N\left(\frac{1}{n}(m_i + m_2 + ... + m_n), \frac{1}{n}\sqrt{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2}\right)$$

Jeżeli zmienne losowe $X_1, X_2, ..., X_n$ są niezależne o takim samym rozkładzie ($X_i - N(m, \sigma)$ dla $i = 1,2,...,n$) to zmienna losowa $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ma rozkład $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

3. Statystyki z próby i ich rozkłady

3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Rozkładem chi – kwadrat (χ^2) nazywamy rozkład zmiennej losowej, która jest sumą kwadratów *n* niezależnych zmiennych losowych $X_1, \ldots, X_n$ o jednakowym rozkładzie normalnym $N(0,1)$.

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

Funkcja gęstości:

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x > 0 \\ 0, & x \le 0 \end{cases}$$

Liczba stopni swobody $\nu = n$.

$E(\chi^2) = n$, $D^2(\chi^2) = 2n$.

Uwaga

Rozkład chi – kwadrat ma **własność addytywności**, tzn. jeśli $\chi_1^2, \ldots, \chi_n^2$ są niezależnymi zmiennymi losowymi o rozkładach chi – kwadrat ze stopniami swobody odpowiednio $\nu_1, \ldots, \nu_n$,to zmienna losowa $\sum_{i=1}^{n} \chi_i^2$ ma rozkład chi – kwadrat o $\nu = \sum_{i=1}^{n} \nu_i$ stopniach swobody.

3. Statystyki z próby i ich rozkłady

3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Rozkładem t – Studenta nazywamy rozkład zmiennej losowej

$$t = \frac{X}{\sqrt{Y}}\sqrt{n}$$

gdzie *X* ma rozkład $N(0,1)$,Y ma rozkład chi – kwadrat z $\nu = n$ stopniami swobody oraz zmienne losowe *X* i *Y* są niezależne.

Funkcja gęstości:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad , \quad t \in \mathbf{R} \quad .$$

Liczba stopni swobody $\nu = n$.

$$E(t) = 0 \quad , \quad D^2(t) = \frac{n}{n-2} \quad , \quad n > 2 .$$

3. Statystyki z próby i ich rozkłady

3.1. Pojęcie statystyki z próby i jej rozkładu, wybrane rozkłady

Rozkładem Fishera – Snedecora (F - Snedecora) nazywamy rozkład zmiennej losowej

$$F = \frac{\chi^2_{n_1} / n_1}{\chi^2_{n_2} / n_2} \quad ,$$

gdzie $\chi^2_{n_1}$ i $\chi^2_{n_2}$ są niezależnymi zmiennymi losowymi o rozkładach chi-kwadrat z n_1 i n_2 stopniami swobody.

Funkcja gęstości:

$$f(z) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_2}{n_1}\right)^{\left(\frac{n_2}{2}\right)} z^{\left(\frac{n_1}{2}-1\right)} \left(z+\frac{n_2}{n_1}\right)^{-\frac{n_1+n_2}{2}} \quad , \qquad z > 0 \qquad .$$

$$E(F) = \frac{n_2}{n_2-2} \quad , \qquad n_2 > 2 \qquad , \quad D^2(F) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)} \qquad , \quad n_2 > 4 \quad .$$

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład średniej arytmetycznej z próby z populacji normalnej

Twierdzenie

Jeśli zmienne losowe $X_1,\ldots,X_n$ są niezależne i o jednakowym rozkładzie normalnym $N(m,\sigma)$, to zmienna losowa

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

ma rozkład normalny ze średnią $E(\overline{X})=m$ i odchyleniem standardowym $D(\overline{X})=\frac{\sigma}{\sqrt{n}}$.

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład średniej arytmetycznej z próby z populacji normalnej z nieznanym odchyleniem standardowym

Twierdzenie

Jeśli $X_1,\ldots,X_n$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(m,\sigma)$, to statystyka

$$t=\frac{\overline{X}-m}{S}\sqrt{n} \qquad \left(t=\frac{\overline{X}-m}{\tilde{S}}\sqrt{n-1}\right) \quad ,$$

gdzie $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i^2$ $\qquad S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ $\qquad \left(\tilde{S}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)$

ma rozkład t – Studenta z $\nu=n-1$ stopni swobody.

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład różnicy średnich arytmetycznych z dwóch prób z populacji normalnych przy znanych odchyleniach standardowych

Niech będą dane dwie populacje o rozkładach normalnych $N(m_1,\sigma_1)$ i $N(m_2,\sigma_2)$, z których pobiera się próby liczące odpowiednio n_1 i n_2 elementów.

Wiemy, że $\overline{X}_1 \sim N\left(m_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$, a $\overline{X}_2 \sim N\left(m_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$.

Wówczas $$\overline{X}_1 - \overline{X}_2 \sim N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład różnicy średnich arytmetycznych z dwóch prób z populacji normalnych przy nieznanych (ale jednakowych) odchyleniach standardowych

Niech będą dane dwie populacje o rozkładach normalnych $N(m_1,\sigma)$ i $N(m_2,\sigma)$, z których pobiera się próby liczące odpowiednio n_1 i n_2 elementów, następnie wyznaczamy średnie $\overline{X}_1$ i $\overline{X}_2$ oraz wariancje S_1^2 i S_2^2 oraz średnia ważoną z obu prób $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.

Wówczas statystyka $t = \frac{(\overline{X}_1-\overline{X}_2)-(m_1-m_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

ma rozkład t – Studenta z $\nu = n_1+n_2-2$ stopniami swobody.

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład wariancji z próby z populacji normalnej

Twierdzenie

Jeśli $X_1,\ldots,X_n$ jest prostą próbą losową z populacji o rozkładzie normalnym, to statystyka

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad \left(\chi^2 = \frac{n\tilde{S}^2}{\sigma^2}\right)$$

ma rozkład chi – kwadrat o $\nu = n-1$ stopniach swobody.

3. Statystyki z próby i ich rozkłady

3.2. Rozkłady dokładne statystyk z próby

Rozkład ilorazu wariancji z prób z dwóch populacji normalnych:

Załóżmy, że z dwóch niezależnych populacji o rozkładzie normalnym z dowolnymi średnimi oraz wariancjami σ_1^2 i σ_2^2 pobiera się próby liczące odpowiednio n_1 i n_2 elementów.

Wówczas statystyka

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

ma rozkład F – Snedecora $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniach swobody.

Wartości rozkładu F-Snedecora są tablicowane, podobnie jak i wcześniej omawiane rozkłady.

3. Statystyki z próby i ich rozkłady
3.3. Rozkłady graniczne statystyk z próby

Z **rozkładów granicznych statystyk z próby** korzysta się, kiedy **rozkład populacji nie jest znany**.

Wyznaczanie rozkładu granicznego nie wymaga na ogół żadnych założeń co do postaci rozkładu populacji generalnej, natomiast wymagana jest **duża liczebność próby**.

3. Statystyki z próby i ich rozkłady

3.3. Rozkłady graniczne statystyk z próby

Niech zmienna X ma rozkład dwumianowy z parametrami n i p . W praktyce korzysta się często ze statystyki $W = \frac{X}{n}$ (będącej częstością sukcesów w n doświadczeniach), która posiada rozkład dwumianowy z parametrami:

E(W) = p oraz $D(W) = \sqrt{\frac{p(1-p)}{n}}$.

Z **twierdzenia Moivre'a-Laplace'a** wynika, że $W = \frac{X}{n}$

ma **graniczny rozkład normalny** $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Czyli, gdy **n jest dostatecznie duże**, można przyjąć, że W posiada w przybliżeniu rozkład normalny.

3. Statystyki z próby i ich rozkłady

3.3. Rozkłady graniczne statystyk z próby

Jeśli zmienne X_1 i X_2 mają rozkłady dwumianowe z parametrami p_1 i p_2, to statystyka postaci:

$$W_1 - W_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

przy $n_1 \to \infty$ i $n_2 \to \infty$ ma **graniczny rozkład normalny** ze średnią

$$E(W_1 - W_2) = p_1 - p_2$$

i odchyleniem standardowym

$$D(W_1 - W_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad .$$

Można wnioskować o tym w oparciu o własność addytywności rozkładu normalnego.

3. Statystyki z próby i ich rozkłady

3.3. Rozkłady graniczne statystyk z próby

Niech zmienna X ma dowolny rozkład ze średnią m i odchyleniem standardowym σ .

Z **twierdzenia Lindeberga-Levy'ego** wynika, że rozkład średniej z próby

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

zmierza przy $n \to \infty$ do rozkładu normalnego z wartością oczekiwaną m i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$,

czyli do rozkładu $N(m; \frac{\sigma}{\sqrt{n}})$.

3. Statystyki z próby i ich rozkłady

3.3. Rozkłady graniczne statystyk z próby

Niech zmienne X_1 i X_2 mają dowolne rozkłady z parametrami, odpowiednio, m_1 i σ_1 , m_2 i σ_2 .

Różnica średnich z próby $\overline{X_1}$ - $\overline{X_2}$

ma przy $n_1 \to \infty$ i $n_2 \to \infty$ **graniczny rozkład normalny** z parametrami

$m_1 - m_2$ i $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$,

czyli rozkład $N(m_1 - m_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

4. Estymatory – własności – metody pozyskiwania
4.1. Podstawowe pojęcia

Estymacja (teoria estymacji) jest podstawowym działem wnioskowania statystycznego.

Stanowi ona zbiór metod pozwalających na wnioskowanie o postaci rozkładu populacji generalnej (tzn. o wartości parametrów rozkładu lub o jego postaci funkcyjnej).

Teoria estymacji wiąże się z nazwiskami: K.Pearsona, R.A. Fishera, J. Neymana.

4. Estymatory – własności – metody pozyskiwania
4.1. Podstawowe pojęcia

Pojęcia, oznaczenia:

θ - parametr populacji generalnej, jest wielkością stałą i jednocześnie nieznaną.

T_n - estymator parametru θ.

Estymatorem T_n parametru θ w populacji generalnej nazywamy statystykę z próby:

$T_n = t(X_1, X_2, \ldots, X_n)$,która służy do oszacowania parametru θ.

Estymator definiowany jako statystyka z próby, jest **zmienną losową** i jako zmienna posiada określony rozkład.

Rozkład estymatora Tn jest uzależniony od rozkładu zmiennej losowej X w populacji generalnej.

4. Estymatory – własności – metody pozyskiwania
4.1. Podstawowe pojęcia

t_n - ocena parametru θ

$t_n = t(x_1, x_2, \ldots, x_n)$ - jest konkretną wartością liczbową, jaką przyjmuje estymator T_n parametru dla realizacji próby $(x_1, x_2, \ldots, x_n)$.

Ocena t_n -jest realizacją zmiennej losowej T_n .

Ocena parametru θ (t_n) jest tą wielkością, jaką w estymacji punktowej przyjmuje się za oszacowanie wartości parametru θ.

4. Estymatory – własności – metody pozyskiwania

4.1. Podstawowe pojęcia

Błędem szacunku (estymacji) parametru θ nazywamy różnicę pomiędzy estymatorem a wartością parametru :

$$T_n - \theta = d$$

błąd szacunku jest zmienną losową a za **miarę tego błędu** przyjmuje się wyrażenie:

$$\Delta = E(T_n - \theta)^2$$

T_n - estymator

D (T_n) - wariancja estymatora T_n

D(T_n) - odchylenie standardowe estymatora T_n, nazywane **średnim błędem (standardowym błędem) szacunku** parametru θ

Wyrażenie $\frac{D(T_n)}{\theta}$ nazywać będziemy **względnym błędem szacunku**.

D(T_n) - powszechnie przyjęło się traktowanie jako podstawowy parametr określający dokładność estymacji punktowej.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

1. **Nieobciążoność** (asymptotyczna nieobciążoność)
2. **Zgodność**
3. **Efektywność** (asymptotyczna efektywność)
4. **Wystarczalność** zwana inaczej dostatecznością

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Nieobciążoność

Definicja 1.

Mówimy, że estymator T_n jest **nieobciążonym estymatorem** parametru θ, jeśli spełniona jest relacja $\boldsymbol{E(T_n) = \theta}$.

Spełnienie tego warunku oznacza, że estymator T_n ma rozkład ze średnią równą wartości szacowanego parametru.

Jeśli szacujemy parametr θ przy pomocy estymatora nieobciążonego, to przy dużej liczbie prób, średnia uzyskanych ocen będzie bliska θ.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Przykład

Do estymacji średniej w populacji N(*m*,δ) wykorzystuje się statystykę $\bar{x} = \frac{1}{n}\sum x_i$.

Można wykazać, że $E(\bar{x}) = E(X)$, co oznacza, że statystyka x jest nieobciążonym estymatorem *E(X)*.

Do estymacji wariancji $D^2(X)$ można wykorzystać statystykę $S^2(x) = \frac{1}{n}\sum (x_i - \bar{x})^2$ statystyką tą jest wariancja z próby.

Do znalezienia $E[S^2(x)]$ odwołujemy się do faktu, że statystyka $\frac{nS^2(x)}{D^2(X)}$ ma rozkład χ^2 z n-1 stopniami swobody stąd $E\left(\frac{nS^2(x)}{D^2(X)}\right) = n-1$; ale

$E\left(\frac{nS^2(x)}{D^2(X)}\right) = \frac{n}{D^2(X)}E[S^2(x)]$; więc $\frac{n}{D^2(X)}E[S^2(x)] = n-1$; stąd

$$E[S^2(x)] = \frac{n-1}{n}D^2(X) = D^2(X) - \frac{1}{n}D^2(X)$$

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Wynika stąd, że $S^2(x)$ jest **obciążonym estymatorem** $D^2(x)$,przy czym obciążenie to wynosi $\frac{1}{n}D^2(X)$.

Oznacza to, że oceny parametru $D^2(x)$ uzyskane w oparciu o $S^2(x)$ są systematycznie obciążone (systematycznie zaniżone). Obciążenie to maleje, jeśli n wzrasta.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Warto zaznaczyć w tym miejscu, że: $S^2(x) = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2$

jest **nieobciążonym estymatorem** parametru $D^2(x)$.

$$E\left(S^2\right) = E\left(\frac{n}{n-1}S^2(x)\right) = \frac{n}{n-1}\frac{n-1}{n}D^2(X) = D^2(X)$$

Jeśli nie jest spełniona równość $E(T_n) = \theta$,

wówczas estymator T_n nazywamy **obciążonym**

a wyrażenie $b(T_n) = E(T_n) - \theta$ nazywamy **obciążeniem estymatora.**

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Nieobciążoność

Definicja 2.

Mówimy, że estymator T_n parametru θ jest **asymptotycznie nieobciążony** jeżeli spełniona jest nierówność:

$$\lim_{n\to\infty} b(T_n) = 0 \qquad \text{lub} \qquad \lim_{n\to\infty} E(T_n) = \theta \;.$$

Estymator $S^2(x)$ wariancji $D^2(x)$ w populacji generalnej posiada własność asymptotycznej nieobciążoności.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Zgodność

Ta własność estymatora wiąże się z **dużą liczebnością próby**.

Definicja 3.

Estymator T_n parametru θ jest **zgodny**, jeżeli spełnia relację dla dowolnego ε > 0 $\lim_{n\to\infty} P\{|T_n - \theta| < \varepsilon\} = 1{,}0$.

Estymator T_n parametru θ jest **zgodny**, jeżeli podlega działaniu prawa wielkich liczb, tzn. jeśli jest stochastycznie zbieżny do szacowanego parametru.

Z prawa wielkich liczb Czebyszewa wynika, że średnia arytmetyczna z próby $\overline{X_n}$ jest zgodnym estymatorem wartości oczekiwanej w populacji generalnej.

$$\lim_{n\to\infty} P\left\{\left|\overline{X_n} - E(X)\right| < \varepsilon\right\} = 1{,}0 \quad ; \quad \text{dla} \quad \varepsilon > 0$$

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

UWAGA! Pomiędzy dwoma omówionymi własnościami (nieobciążonością i zgodnością) zachodzą następujące związki:

1. Jeśli estymator T_n parametru θ jest **zgodny to jest asymptotycznie nieobciążony.**

Twierdzenie odwrotne nie jest prawdziwe.

2. Jeśli estymator T_n parametru θ jest **nieobciążony** lub **asymptotycznie nieobciążony** oraz jego wariancja spełnia $\lim_{n\to\infty} D^2(T_n)=0$ to T_n jest estymatorem zgodnym.

Ostatnie twierdzenie jest użyteczne w sprawdzaniu zgodności estymatora

np.: $\overline{X_n}$ estymator *E(X)*, jeśli $D^2\left(\overline{X}\right)=\frac{D^2(X)}{n}$; to $D^2\left(\overline{X_n}\right)\to 0$; jeśli $n\to\infty$.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Efektywność

Definicja 4.

Estymator T_n parametru θ jest **najefektywniejszy**, jeśli wśród estymatorów nieobciążonych ma najmniejszą wariancję.

Jeśli dany jest zbiór wszystkich nieobciążonych estymatorów $T_n^1, T_n^2, \ldots, T_n^r$ parametru θ, to estymator , który ma w tym zbiorze najmniejszą wariancję, tzn.:

$$D^2\left(T_n^*\right) \leq D^2\left(T_n^r\right) \quad , \quad \text{i=1,2...r,}$$

nazywamy **najefektywniejszym estymatorem** parametru θ.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Miarą efektywności estymatora danego jest stosunek wariancji estymatora najefektywniejszego do wariancji estymatora danego, co można zapisać:

$$e\left(T_n^i\right)=\frac{D^2\left(T_n^*\right)}{D^2\left(T_n^i\right)}$$

wyrażenie to określa się **efektywnością estymatora**.

Efektywność estymatora najefektywniejszego jest równa jedności, w pozostałych wypadkach 0< e <1.

Przy określaniu efektywności estymatora należałoby znać wszystkie estymatory nieobciążone szacowanego parametru i ich wariancję lub wiedzieć czemu jest równa wariancja estymatora najefektywniejszego.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

W celu wyznaczeniu wariancji estymatora najefektywniejszego, można skorzystać z **twierdzenia zwanego nierównością RAO-CRAMERA**, które mówi,

że przy pewnych ogólnych warunkach wariancja $D^2(T_n)$ dowolnego nieobciążonego estymatora parametru θ spełnia relację:

$$D^2\left(T_n\right) \geq \frac{1}{nE\left[\frac{\partial \log f\left(x,\theta\right)}{\partial\theta}\right]^2} = D^2\left(T_n^*\right)$$

gdzie *f(x, θ)* oznacza funkcję gęstości lub funkcję prawdopodobieństwa rozkładu populacji generalnej.

Dla określenia wagi estymatora najefektywniejszego, niezbędna jest znajomość postaci rozkładu populacji generalnej.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Asymptotyczna efektywność

Definicja 5.

Mówimy, że estymator T_n parametru θ jest **asymptotycznie**

najefektywniejszy, jeśli

$$\lim_{n\to\infty} e(T_n) = 1{,}0$$

tzn. jeżeli liczebność próby dąży do nieskończoności, wariancja $D^2(T_n)$

estymatora T_n przyjmuje wartości coraz bliższe wartości $D^2(T_n)$

najefektywniejszego estymatora.

4. Estymatory – własności – metody pozyskiwania
4.2. Podstawowe własności estymatorów

Wystarczalność zwana dostatecznością

Mówimy, że estymator T_n parametru θ jest **wystarczający** (dostateczny), jeżeli zawiera wszystkie informacje jakie na temat tego parametru θ występują w próbie, i żaden inny estymator nie może dać dodatkowych informacji o szacowanym parametrze.

Estymator dostateczny nie zawsze istnieje.

Estymator dostateczny, to taki, który dostarcza najwięcej informacji o danym parametrze wśród wszystkich możliwych estymatorów tego parametru.

Inaczej estymator wystarczający jest tak zbudowany, że żaden inny estymator nie może dostarczyć więcej informacji o szacowanym parametrze θ.

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Do najczęściej stosowanych metod należą:

1. ***metoda momentów (MM)***
2. ***metoda największej wiarygodności (MNW)***
3. ***metoda najmniejszych kwadratów (MNK)***
4. ***inne metody (metoda minimalnej straty, metoda najmniejszej odległości itp.)***

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda momentów (metoda analogii)

Najstarsza, najprostsza metoda szacowania

Idea tej metody polega na tym, że momenty (zwykłe lub centralne) można przedstawić jako pewne funkcje parametrów rozpatrywanego rozkładu.

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda momentów

Niech $X_1,\ldots,X_n$ będzie próbą losową.

Momenty z próby rzędu *k*:

zwykłe: $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

centralne: $M'_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - M_1\right)^k$

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda momentów

Niech $X_1,\ldots,X_n$ będzie próbą losową.

Momenty rozkładu zmiennej losowej skokowej i ciągłej rzędu *k*:

zwykłe:

$$m_k = E\left(X^k\right) = \begin{cases} \sum_i x_i^k p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} \mathrm{x}^k f(x)dx & \text{dla zmiennej losowej ciaglej} \end{cases}$$

centralne:

$$\mu_k = E\left[X - E(X)\right]^k = \begin{cases} \sum_i \left[x_i - E(X)\right]^k p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} \left[x - E(X)\right]^k f(x)dx & \text{dla zmiennej losowej ciaglej} \end{cases}$$

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

UWAGA!

Metoda momentów dostarcza estymatorów:
obciążonych i charakteryzujących się niewielką efektywnością.

Wyjątek stanowi tu średnia arytmetyczna jako estymator wartości oczekiwanej, która bez względu na rozkład zmiennej losowej w populacji generalnej ma wszystkie pożądane własności dobrego estymatora, czyli: **zgodność, nieobciążoność, dostateczność.**

A jeśli zmienna losowa X ma rozkład N(m,δ) to średnia arytmetyczna jest również ***najefektywniejszym*** estymatorem wartości oczekiwanej *m*.

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda największej wiarygodności

- Koncepcja sformułowana przez R.A. Fishera w latach 20 tych XX wieku
- Jedna z najbardziej rozpowszechnionych metod estymacji
- Punkt wyjścia - określenie funkcji wiarogodności

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda największej wiarygodności

W procedurze estymacji MNW można wydzielić następujące etapy:

1) określenie funkcji wiarygodności L

2) wyznaczenie logarytmu naturalnego $\ln L$ z tej funkcji

3) wyznaczenie pochodnych cząstkowych względem nieznanych parametrów $\frac{\partial \ln L}{\partial \theta_i}$ dla $i = 1, 2, 3,..., r$

4) rozwiązanie układu równań $\frac{\partial \ln L}{\partial \theta_i} = 0$ względem (dla $i = 1, 2, 3,..., r$).

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda największej wiarygodności

Funkcję wiarygodności buduje się dla rozkładów skokowych lub ciągłych.

W przypadku rozkładów dla zmiennej ciągłej funkcja ma postać:

$$L = (t_1, t_2, t_3, \ldots, t_n, \theta_1, \theta_2, \theta_3, \ldots, \theta_r) = \prod_{i=1}^{n} f(t_i, \theta_1, \theta_2, \theta_3, \ldots, \theta_r)$$

gdzie:

f – funkcja gęstości rozkładu,

$\theta_1, \theta_2, \theta_3, \ldots, \theta_r$ – nieznane parametry.

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Estymatory uzyskane **metodą największej wiarogodności** nie zawsze są nieobciążone, ale mają szereg innych własności.

Przy pewnych ogólnych założeniach estymatory MNW parametru θ charakteryzują się następującymi własnościami:

1. są zgodne,
2. mają asymptotyczny rozkład normalny o wartości oczekiwanej θ i wariancji równej: $$\frac{1}{nE\left[\frac{\partial \log f(x,\theta)}{\partial \theta}\right]^2}$$
3. są one co najmniej asymptotycznie nieobciążone i asymptotycznie najefektywniejsze,
4. jeśli istnieje dostateczny estymator parametru θ to jest on estymatorem MNW,
5. jeśli istnieje najefektywniejszy estymator parametru θ to jest on uzyskany metodą największej wiarogodności MNW.

4. Estymatory – własności – metody pozyskiwania
4.3. Metody pozyskiwania estymatorów

Metoda najmniejszych kwadratów

- Za twórców uważa się A.M. Lagendre'a, K.F. Gaussa, A.A. Markowa.

- Ten typ estymatorów tzn. estymatorów uzyskanych MNK ma szczególne zastosowanie w analizie regresji.

4. Estymatory – własności – metody pozyskiwania
4.4. Przykład wyznaczania estymatora metodą MNW

Przykład wyznaczania estymatora parametru dla rozkładu wykładniczego bez zmiennych.

Funkcja gęstości w rozkładzie wykładniczym jest postaci

$$f(t) = \alpha \mathrm{e}^{-\alpha t} \qquad \text{dla} \qquad \alpha > t,\ t \geq 0 \quad .$$

Kolejne etapy estymacji to:

1) określenie funkcji wiarogodności – L

$$L = L(t_1, t_2, t_3, \ldots, t_n : \alpha) = \alpha^n \prod_{i=1}^{n} \mathrm{e}^{-\alpha t_i}$$

2) wyznaczanie logarytmu naturalnego z tej funkcji – ln$L$

$$\ln L = n \ln \alpha - \alpha \sum t_i$$

4. Estymatory – własności – metody pozyskiwania
4.4. Przykład wyznaczania estymatora metodą MNW

3. wyznaczanie pochodnej cząstkowej względem nieznanego parametru α $\frac{\partial \ln L}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} t_i$
4. rozwiązanie układu równań

$$\frac{\partial \ln L}{\partial \alpha} = 0$$

$$\frac{n}{\alpha} - \sum_{i=1}^{n} t_i = 0 \qquad \frac{n}{\alpha} = \sum_{i=1}^{n} t_i \Big| : n$$

$$\frac{1}{\hat{\alpha}} = \frac{1}{n}\sum_{i=1}^{n} t_i \qquad \frac{1}{\hat{\alpha}} = \bar{t}$$

$$\hat{\alpha} = \frac{1}{\bar{t}}$$ – estymator parametru α

Dodatkowo: Idea podejścia bayesowskiego

W podejściu bayesowskim parametry modelu traktowane są jak zmienne losowe.

Wnioskowanie o nieznanych parametrach bazuje na ich rozkładach a posteriori uzyskanych przy pomocy twierdzenia Bayesa poprzez łączenie informacji z rozkładu a priori i dostępnych danych.

Rozkład a priori umożliwia włączenie posiadanej wiedzy dotyczącej prawdopodobnego zakresu wartości estymowanych parametrów.

Jeśli takiej wiedzy nie posiadamy, to możemy wykorzystać nieinformacyjny rozkład a priori, wówczas wyniki analizy bayesowskiej uzyskane tą metodą będą bardzo podobne do wyników uzyskanych metodą klasyczną bazującą na funkcji wiarogodności.

Uzyskanie rozkładu a posteriori wymaga często wykorzystania metod symulacji MCMC,

Załóżmy, że jesteś zainteresowany w oszaciowaniu θ z danych y = {y1, ..., yn} poprzez użycie modelu statystycznego opisanego przez gęstość p(y|θ). Podejście Bayesowskie twierdzi, że θ nie może być określone dokładnie, a niepewność dotycząca parametrów jest wyrażona poprzez prawdopodobieństwo i rozkład. Można powiedzieć, że θ o normalnym rozkładzie ze średnią 0 i wariancją 1, jeśli podejrzewa się, że ten rozkład najlepiej opisuje niepewność związaną z parametrem.

Następujące kroki opisują istotne elementy wnioskowania Bayesowskiego:

1. Rozkład prawdopodobieństwa dla θ jest sformułowany jako $\pi(\theta)$, co znane jest jako rozkład a priori. Rozkład a priori wyraża twoje przekonania dotyczące np. średniej, rozstępu czy skośności parametru zanim przeanalizujesz dane.
2. Mając dane y , wybierasz model statystyczny $p(y|\theta)$, by opisywał rozkład y danego .
3. Uaktualniasz swoje przekonania dotyczące θ poprzez połączenie informacji z rozkładu a priori i z danych poprzez policzenie rozkładu a posteriori $p(\theta|y)$.

Trzeci krok przeprowadza się używając teorii Bayesa, która pozwala połączyć rozkład a priori i model w następujący sposób:

$$p(\theta|y) = p(\theta, y)/p(y) = p(y|\theta)\,\pi(\theta)/p(y) = p(y|\theta)\,\pi(\theta) / \int p(y|\theta)\,\pi(\theta)\,d\theta$$

Wielkość

$$p(y) = \int p(y|\theta)\,\pi(\theta)\,d\theta$$

jest stałą normalizującą rozkładu a posteriori. p(y) jest również rozkładem krańcowym y i jest czasem nazywana rozkładem normalnym danych.

Funkcja prawdopodobieństwa θ jest dowolną funkcją proporcjonalną do $p(y|\theta)$ – tj. $L(\theta) \propto p(y|\theta)$. Innym sposobem na zapisanie teorii Bayesa jest

$$p(\theta|y) = L(\theta)\,\pi(\theta) / \int L(\theta)\,\pi(\theta)\,d\theta$$

Rozkład krańcowy p(y) jest całkowity; stąd, tak długo jak jest skończony, konkretna wartość liczby całkowitej nie dostarcza żadnych dodatkowych informacji o rozkładzie a posteriori. Dlatego też $p(\theta|y)$ może być przypisana do dowolnej stałej, zaprezentowanej tu w formie proporcjonalnej jako

$$p(\theta|y) \propto L(\theta)\,\pi(\theta)$$

Mówiąc prosto, teoria Bayesa mówi ci jak aktualizować posiadaną wiedzę na podstawie nowych informacji. Zaczynasz z przekonaniem a priori $\pi(\theta)$ i, po zdobyciu informacji z danych y, zmieniasz lub aktualizujesz swoje przekonania o θ i otrzymujesz $p(\theta|y)$. To są najistotniejsze elementy podejścia Bayesowskiego do analizy danych .

DZIĘKUJĘ ZA UWAGĘ!

DZIĘKUJĘ ZA UWAGĘ!