PODSTAWY BIOSTATYSTYKI dla ZM III

dr inż Krzysztof Bryś

Wykład 2

Podstawowe teoretyczne rozkłady prawdopodobieństwa zmiennej losowej jednowymiarowej

Typu skokowego

1. Rozkład jednopunktowy.

Funkcja prawdopodobieństwa : $P(X = c) = 1$ dla pewnej stałej c
Wartość oczekiwana: $E(X) = c$
Wariancja: $D^2(X) = 0$
Interpretacja: Rozkład dowolnej stałej liczbowej X.

2. Rozkład dwupunktowy (zerojedynkowy).

Funkcja prawdopodobieństwa : $P(X = 1) = p$, $P(X = 0) = q = 1 - p$
Wartość oczekiwana: $E(X) = p$
Wariancja: $D^2(X) = p \cdot q = p \cdot (1 - p)$
Interpretacja: Rozkład dowolnej zmiennej X, która odpowiada na pewne pytanie albo TAK (X = 1-"sukces") albo NIE (X = 0-"porażka"), rozkład dowolnej cechy "zero-jedynkowej" (obiekt albo ją posiada ($X = 1$) albo nie posiada ($X = 0$).

3. Rozkład Bernoulliego (dwumianowy) - $B(n, p)$

Schemat doświadczeń Bernoulliego:
- n niezależnych doświadczeń,
- w każdym doświadczeniu albo sukces z prawdopodobieństwem p albo porażka (z prawdopodobieństwem $q = 1 - p$);

Interpretacja: Zmienna losowa X ma rozkład $B(n, p)$ jeśli mówi o liczbie sukcesów w schemacie n niezależnych doświadczeń Bernoulliego z prawdopodobieństwem sukcesu p w każdym z nich. Jest sumą n niezależnych zmiennych losowych o rozkładzie zerojedynkowym.
Funkcja prawdopodobieństwa : $P(X = k) = \binom{n}{k} p^k \cdot q^{n-k}$ dla $k = 0, 1, 2, \ldots, n$, $q = 1 - p$.
Wartość oczekiwana: $E(X) = np$
Wariancja: $D^2(X) = n \cdot p \cdot q$

4. Rozkład Poissona - $\mathcal{P}o(\lambda)$

Funkcja prawdopodobieństwa : $P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$ dla $k = 0, 1, 2, \ldots$
Wartość oczekiwana: $E(X) = \lambda$
Wariancja: $D^2(X) = \lambda$
Interpretacja: Rozkład graniczny dla rozkłaadu $B(n, p)$ przy $n \to +\infty$.
Dla dostatecznie dużych n, zmienna losowa o rozkładzie $B(n, p)$ ma w przybliżeniu rozkład Poissona z parametrem $\lambda = n \cdot p$.

Typu ciągłego

1. Rozkład jednostajny na przedziale $(a; b)$ - $U(a, b)$

Funkcja gęstości prawdopodobieństwa :

$$f(x) = \begin{cases} \frac{1}{b-a} & , \text{ dla } a < x < b \\ 0 & , \text{ dla pozostałych x} \end{cases}$$

Wartość oczekiwana: $E(X) = \frac{a+b}{2}$
Wariancja: $D^2(X) = \frac{(b-a)^2}{12}$
Interpretacja Zmienna losowa X ma rozkład $U(a, b)$ jeśli przyjęcie przez tą zmienną dowolnej wartości z przedziału $(a; b)$ jest jednakowo prawdopodobne.

2. Rozkład normalny (Gaussa) - $N(m, \sigma)$

Funkcja gęstości prawdopodobieństwa : $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$ dla $x \in R$
Wartość oczekiwana: $E(X) = m$
Wariancja: $D^2(X) = \sigma^2$
Wykresem powyższej funkcji gęstości prawdopodobieństwa jest **krzywa Gaussa**
Zmienna losowa standaryzowa dla zmiennej losowej o rozkładzie $N(m, \sigma)$:

$$\overline{X} = \frac{X - m}{\sigma}$$

ma rozkład normalny standardowy $N(0, 1)$.

Dystrybuanta rozkładu normalnego standardowego $N(0, 1)$**:**

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} dt \text{ dla } x \in R$$

Z parzystości funkcji gęstości prawdopodobieństwa rozkładu $N(0, 1)$ wynika, że:

$$\Phi(-x) = 1 - \Phi(x).$$

u_α - kwantyl rzędu α zmiennej losowej o rozkładzie $N(0, 1)$ (tzn. $\Phi(u_\alpha) = \alpha$)

3. Rozkład chi kwadrat o n stopniach swobody

Zmienna losowa $\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2$, gdzie $X_1, X_2, \ldots X_n$ zmienne o rozkładzie $N(0, 1)$ ma rozkład chi-kwadrat o n stopniach swobody
Wartość oczekiwana: $E(\chi^2) = n$
Wariancja: $D^2(\chi^2) = 2n$

Dla dużych n ($n > 40$) rozkład chi-kwadrat o n stopniach swobody można przybliżać rozkładem $N(n, \sqrt{2n})$.

$\chi^2(\alpha, n)$ = kwantyl rzędu $1 - \alpha$ zmiennej o rozkładzie chi-kwadrat o n stopniach swobody

4. Rozkład t-Studenta o n stopniach swobody.

Zmienna losowa $T = \frac{X}{\sqrt{\frac{\chi^2}{n}}}$, gdzie X zmienna losowa o rozkładzie $N(0, 1)$ a zmienna χ^2 ma rozkład chi-kwadrat o n stopniach swobody.
Wartość oczekiwana: $E(T) = 0$.
Wariancja: $D^2(T) = \frac{n}{n-2}$.

Dla dużych n ($n > 40$) rozkład t-Studenta o n stopniach swobody można przybliżać rozkładem $N(0, 1)$.

$t(\alpha, n)$ = kwantyl rzędu $1 - \frac{\alpha}{2}$ zmiennej o rozkładzie t-Studenta o n stopniach swobody.

Statystyka - pojęcia wstępne

populacja - cały zbiór badanych przedmiotów lub wartości.
próba - skończony podzbiór populacji podlegający badaniu.
próba losowa - próba losowana (najczęściej) zgodnie z rozkładem równomiernym, tzn. wylosowanie każdej próby jest jednakowo prawdopodobne.
cechy: mierzalne, niemierzalne
badana cecha = zmienna losowa X
Poszukiwany: rozkład cechy w populacji = rozkład zmiennej losowej X
próba n-elementowa = ciąg n niezależnych zmiennych losowych $(X_1, \ldots, X_n)$ o jednakowym rozkładzie (takim jak poszukiwany rozkład zmiennej losowej X).

Etapy badania statystycznego

1) Przygotowanie (formatowanie) badania (określenie celu, rodzaju, potrzebnych parametrów wejściowych badania).

2) Przeprowadzenie badania (wylosowanie próby i określenie wartości badanych cech w próbie).
3) Zebranie uzyskanych podczas badania danych.
4) Opis i wnioskowanie statystyczne (obliczenie parametrów, estymacja, weryfikacja hipotez).
5) Przedstawienie wyników.

Szeregi statystyczne

1) **Szereg wyliczający uporządkowany:** $(x_1, x_2, \ldots, x_n)$
przy czym $x_1 \leq x_2 \leq \ldots \leq x_n$.
2) **Szereg rozdzielczy punktowy:** $(x_1, x_2, \ldots, x_k)$, $(n_1, n_2, \ldots, n_k)$,
gdzie $x_1 < x_2 < \ldots < x_k$ oraz dla każdego $i = 1, 2, \ldots, k$: n_i-liczba realizacji (obserwacji) wartości x_i, $\sum_{i=1}^{k} n_i = n$.
3) **Szereg rozdzielczy przedziałowy:** $(y_0; y_1 >, (y_1; y_2 >, \ldots, (y_{k-1}; y_k)$, $(n_1, n_2, \ldots, n_k)$,
gdzie $y_0 < y_1 < y_2 < \ldots < y_{k-1} < y_k$ oraz dla każdego $i = 1, 2, \ldots, k$: n_i-liczba realizacji (obserwacji) wartości należącej do przedziału $(y_{i-1}; y_i)$, $\sum_{i=1}^{k} n_i = n$.
Wszystkie wartości należące do przedziału $(y_{i-1}; y_i >$, $i = 1, 2, \ldots, k$ utożsamia się z jego środkiem x_i.
Reguły wyznaczania liczby przedziałów (klas): $k \approx \sqrt{n}$, $k \leq 5 \log n$.

Parametry empiryczne

Miary położenia rozkładu

1) **Średnia z próby** $\overline{x}$
- dla szeregu wyliczającego:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- dla szeregu rozdzielczego:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot x_i$$

2) **Dominanta (moda, wartość modalna)** D = punkt, w którym funkcja prawdopodobieństwa osiąga największą wartość
- dla szeregu wyliczającego: najczęściej występująca wartość,
- dla szeregu rozdzielczego punktowego: punkt, dla którego liczebność (częstość) osiąga największą wartość, - dla szeregu rozdzielczego przedziałowego (wzór interpolacyjny):

$$D = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot h_d,$$

gdzie
x_{0d} - początek przedziału zawierającego dominantę (przedziału o najwiekszej liczebności),
h_d - szerokość przedziału zawierającego dominantę (przedziału o najwiekszej liczebności),
n_d - liczebność przedziału zawierającego dominantę (najwieksza liczebność),
n_{d-1} - liczebność przedziału poprzedzającego przedział zawierający dominantę,
n_{d+1} - liczebność przedziału następnego po przedziale zawierającym dominantę.
3) **Dystrybuanta empiryczna (częstość skumulowana** $F_n(x)$
- dla szeregu wyliczającego:

$$F_n(x) = \frac{1}{n} |\{i : x_i < x, i = 1, \ldots, n\}|$$

- dla szeregu rozdzielczego:

$$F_n(x) = \sum_{i: x_i < x} \frac{n_i}{n}$$

4) **Kwantyl empiryczny rzędu** p $x_{p,n}$:
(punkt w którym dystrybuanta empiryczna po raz pierwszy osiąga wartość niemniejszą niż p)
- dla szeregu wyliczającego:

$$x_{p,n} = x_{\lceil np \rceil}$$

- dla szeregu rozdzielczego punktowego:

$$x_{p,n} = x_q \text{ gdzie } q = \min\{r : p \leq \sum_{i=1}^{r} \frac{n_i}{n}\}$$

- dla szeregu rozdzielczego przedziałowego (wzór interpolacyjny):

$$x_{p,n} = x_{0p} + (np - \sum_{x_i < x_{0p}} n_i) \cdot \frac{h_p}{n_p},$$

gdzie
x_{0p} - początek przedziału zawierającego $x_{p,n}$ (przedziału w którym dystrybuanta empiryczna po raz pierwszy osiąga wartość niemniejszą niż p),
h_p -szerokość przedziału zawierającego $x_{p,n}$,
n_p -liczebność przedziału zawierającego $x_{p,n}$,
$\sum_{x_i<x_{0p}} n_i$ - liczebność skumulowana dla przedziału poprzedzającego przedział zawierający $x_{p,n}$ (suma liczebności przedziałów poprzedzających)
Mediana: Me = kwantyl rzędu $\frac{1}{2}$
Kwartyl dolny: Q_1 = kwantyl rzędu $\frac{1}{4}$
Kwartyl górny: Q_3 = kwantyl rzędu $\frac{3}{4}$.

Miary rozproszenia rozkładu

5) **Wariancja z próby** s^2
- dla szeregu wyliczającego:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- dla szeregu rozdzielczego:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^2$$

6) **Odchylenie standardowe z próby** $s = \sqrt{s^2}$.

7) **Współczynnik zmienności** $V = \frac{s}{\overline{x}} \cdot 100\%$.

8) **Rozstęp** R = różnica między największą i najmniejszą wartością w próbie.

9) **Współczynnik asymetrii** A_s:
- dla szeregu wyliczającego:

$$A_s = \frac{1}{s^3} \cdot (\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3)$$

- dla szeregu rozdzielczego:

$$A_s = \frac{1}{s^3} \cdot (\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^3)$$

10) **Kurtoza (współczynnik skupienia)** A_s:
- dla szeregu wyliczającego:

$$K = \frac{1}{s^4} \cdot (\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4)$$

- dla szeregu rozdzielczego:

$$K = \frac{1}{s^4} \cdot (\frac{1}{n}\sum_{i=1}^{k} n_i \cdot (x_i - \overline{x})^4)$$

11) **Współczynnik skośności** A_1:

$$A_1 = \frac{\overline{x} - D}{s}$$