Rachunek Prawdopodobieństwa

Prof. dr hab. Mieczysław Wodecki

Zasady zaliczeń

Forma zaliczenia – zaliczenie na ocenę (kolokwium - na ostatnim wykładzie).

Termin poprawkowy – w sesji.

Rachunek Prawdopodobieństwa*(w. 15)

1. **Elementy statystyki – miary położenia, rozproszenia**
2. **Przestrzeń zdarzeń elementarnych. Własności prawdopodobieństwa. Prawdopodobieństwo klasyczne i geometryczne**
3. **Prawdopodobieństwo warunkowe. Wzór Bayesa. Niezależność zdarzeń.**
4. **Zmienne losowe (rozkłady: dwumianowy, Poissona, normalny, wykładniczy).**
5. **Parametry zmiennych losowych, standaryzacja, tablice.**
6. **Wartości oczekiwane, wariancje, mediany i kwartale wybranych rozkładów.**
7. **Zmienne losowe dwuwymiarowe. Rozkłady brzegowe. Niezależność zmiennych losowych. Współczynnik korelacji.**
8. **Ciągi zmiennych losowych: sumowanie niezależnych zmiennych losowych. Prawo wielkich liczb (słabe).**

Literatura

1. **H. Jasiulewicz, W. Kordecki, Rachunek prawdopodobieństwa i statystyka matematyczna.**
2. **A. Plucińska, E. Pluciński, Probabilistyka, WNT, Warszawa 2006.**

** dokładny program oraz pełny spis literatury są zamieszczone w Sylabusie.*

Rachunek prawdopodobieństwa (inaczej probabilistyka) jest działem matematyki, który zajmuje się modelowaniem zjawisk (zdarzeń) przypadkowych (losowych). Czyli takich, których wyniku zakończenia nie da się w sposób jednoznaczny przedstawić.

Poszukuje się prawidłowości dotyczących możliwości uzyskania oczekiwanego wyniku zdarzeń losowych. Powstanie tego działu matematyki (XVII wiek) było związane z hazardem i poznania szans na wygranie w konkretnej grze losowej.

Dziś osiągnięcia tej dziedziny nauki wykorzystuje się w wielu innych dziedzinach, takich jak nauki przyrodnicze,społecznych i techniczne.

Elementy rachunku prawdopodobieństwa są stosowane, między innymi w:

- **metodach Monte Carlo,**
- **sieciach neuronowych,**
- **uczeniu maszynowym,**
- **różnego rodzaju zadaniach optymalizacji i klasyfikacji.**

Elementy probabilistyki są także stosowane w informatyce:

- **modelowanie procesów z niepewnymi (np. losowymi) parametrami,**
- **algorytmy zrandomizowane (losowy wybór w trakcie działanie), np. algorytm: genetyczny, mrówkowy, symulowanego wyżarzania, itd.,**
- **analiza probabilistyczna algorytmu („zachowanie” algorytmu na losowych danych),**
- **losowe eksperymenty (porównywanie wyników algorytmów – metody statystyki opisowej).**

Opracowanie danych statystycznych

Statystyka zajmuje się zbieraniem, opracowywaniem i analizą danych o zjawiskach masowych (nieprzewidywalnych), tj. zjawiskach powtarzających się dużą liczbę razy.

Badanie statystyczne **umożliwia ustalenie pewnych prawidłowości (regularności), charakterystycznych wartości oraz tendencji w obserwowanych procesach. Nie można ich zazwyczaj ustalić na podstawie pojedynczych obserwacji.**

Badaniu podlega populacja (zbiorowość statystyczna) – zbiór elementów (jednostek statystycznych) mających wspólną własność (cechę).

Cecha może przyjmować wartości:

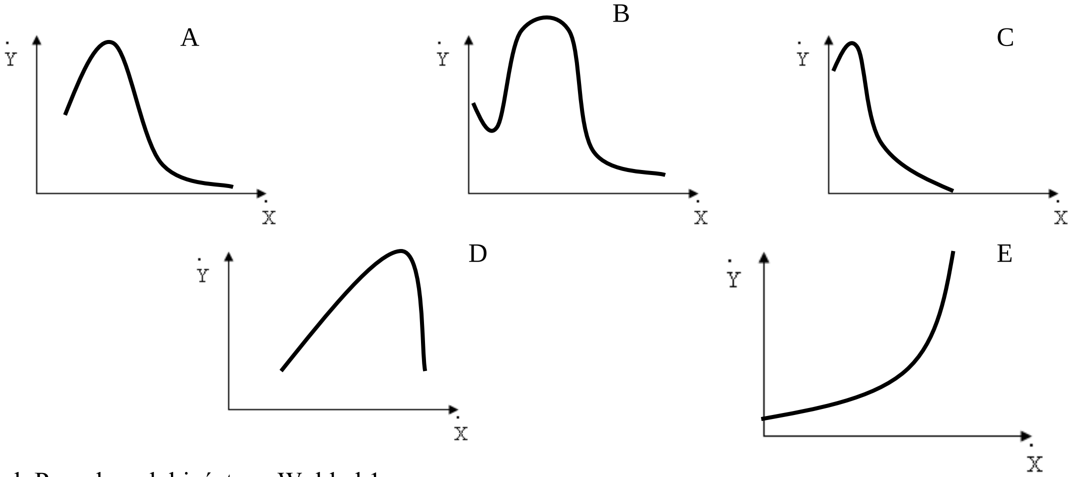
a) ***dyskretne*** **(skokowe) – np. liczba dzieci w rodzinie, liczba braków, liczba wypadków,**
b) ***ciągłe*** **(z pewnego przedziału) – np. zużycie wody, energii.**

Badanie statystyczne można przeprowadzić:

1. **poddając obserwacji wszystkie elementy populacji,**
2. **poddać obserwacji pewną reprezentację (próbę z populacji) i na tej podstawie wyciągnąć wnioski dotyczące całej populacji.**

   Ponieważ populacje bywają bardzo duże, stąd zazwyczaj stosuje się 2-gą metodę.

Rozkłady cechy w populacji



<u>Statystyczne Opracowanie Wyników</u>

Niech X będzie badaną cechą w pewnej populacji.

$x_1, x_2, \dots, x_n$ ciąg *n* wartości obserwacji cechy *X* (wartości próby) nazywamy *szeregiem statystycznym (szczegółowym)*.

Po uporządkowaniu (rosnąco lub malejąco) otrzymujemy *uporządkowany szereg statystyczny*.

Charakterystyki liczbowe rozkładu cechy w populacji

Analiza danych statystycznych powinna doprowadzić do zwięzłego przedstawienia wyników. Używa się do tego odpowiednich charakterystyk (liczb) zwanych *parametrami statystycznymi*(parametrami opisu danych statystycznych). Parametry te dzieli się na:

1. **miary położenia,**
2. **miary zmienności (rozproszenia),**
3. **miary asymetrii,**
4. **miary koncentracji.**

Miary położenia

Średnia arytmetyczna

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

Dominanta **(modalna, moda)**

Wielkość D najliczniej występująca w szeregu.

Dla szeregu: 18, 21, 27, 29, 35, 43, 51 dominanta nie istnieje.

Dla szeregu: 8, 1, 2, 2, 3, 1, 4, 1, 7 D=1.

Mediana ***(„wartość środkowa")***

Wielkość M taka, że co najmniej 50% obserwacji nie przekracza M ($\leq$) oraz co najmniej 50% obserwacji jest nie mniejsza od M ($\geq$).

Dla szeregu: 18, 21, 27, 29, 35, 43, 51 M=29.
Dla szeregu: 18, 21, 27, 29, 35, 43, 51, 59
M=(29+35)/2=32

Kwantyle

Kwantyl$p$-procentowy, to wielkość K_p taka, że co najmniej p% obserwacji nie przekracza K_p ($\leq$) oraz co najmniej (100-p)% obserwacji jest nie mniejsza od K_p($\geq$).

3,4,6,8,13,15,16,20,21,23,25,33,44,51,52,56

$$\mathrm{K}_{25\%} = \frac{8+13}{2} = 10{,}5$$

Miary zmienności (zróżnicowania, rozproszenia)

Wariancja *(oznaczana przez s^2, σ^2, D^2)*

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

Dla Przykładu 1 $18,\ 21,\ 27,\ 29,\ 35,\ 43,\ 51,\quad \overline{x} = 32.$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{7}[(18-32)^2 + (21-32)^2 + (27-32)^2 + (29-32)^2 +$$

$$+(35-32)^2 + (43-32)^2 + (51-32)^2] = 120.2857$$

Odchylenie standardowe (oznaczana przez s, σ, D)

Jest to pierwiastek kwadratowy z wariancji.

Przedział $[\overline{x} - s, \overline{x} + s]$ nazywamy typowym przedziałem zmienności. Należą do niego niemal wszystkie elementy 70% (dużo)

Współczynnik zmienności

$$V(s) = \frac{s}{\overline{x}} \bullet 100\%$$

umożliwia porównywanie zmienności różnych cech.

Szereg rozdzielczy

Najczęściej ze względu na dużą liczbę wartości cechy w szeregu statystycznym, tworzymy szereg rozdzielczy, w których klasami są przedziały.

Dany jest szereg statystyczny $x_1, x_2, \dots, x_n$. **Niech**

$$x_{\min} = \min\{x_1, x_2, \dots, x_n\} \quad \text{oraz} \quad x_{\max} = \max\{x_1, x_2, \dots, x_n\}.$$

Różnicę $R = x_{\max} - x_{\min}$ **nazywamy *rozstępem* szeregu.**

Przyjmuje się, że liczba klas $k \approx \sqrt{n}$ **i zaleca się, aby**

Liczba obserwacji n	Liczba klas k
40-60	6-8
60-100	7-10
100-200	9-12
200-500	11-17

Klasy są rozłącznymi przedziałami $[x_{0,i},\ x_{1,i})$, $i = 1, 2, \dots, k$, **gdzie** $x_{0,i}$ **oraz** $x_{1,i}$ **są odpowiednio *dolną* i *górną granicą* klasy.**

Różnicę $h = x_{1,i} - x_{0,i}$ **nazywamy *rozpiętością* klasy,**

$$h \approx \frac{x_{\max} - x_{\min}}{k} = \frac{R}{k}.$$

Przykład 1.

Spytano o wiek 20 losowo wybranych pracowników pewnego zakładu.

19, 22, 24, 24, 25, 25, 26, 26, 27, 29, 30, 30, 31, 32, 32, 34, 34, 35, 35, 35

$$n=20, \quad k \approx \sqrt{n} = \sqrt{20} \approx 4.$$

$$x_{\min} = \min\{x_1, x_2, \dots, x_{20}\} = 19,$$

$$x_{\max} = \max\{x_1, x_2, \dots, x_{20}\} = 35.$$

$$R = x_{\max} - x_{\min} = 35 - 19 = 16. \; h \approx \frac{x_{\max} - x_{\min}}{k} = \frac{R}{k} = \frac{16}{4} = 4.$$

Mamy więc klasy: $[19, 23),\ [23, 27),\ [27, 31),\ [31, 35)$.

Tabela 1. Szereg rozdzielczy przedziałowy.

numer klasy	Klasa $[x_{0,i}, x_{1,i})$	liczebność n_i	Częstość n_i/n ω_i	liczebność skumulowana $n_{sk,i}$	częstość skumulowana $\omega_{sk,i}$

1	19-23	2	0,1	2	0,1
2	23-27	6	0,3	8	0,4
3	27-31	4	0,2	12	0,6
4	31-35	8	0,4	20	1,0

UWAGI

1. **Przedstawienie w postaci szeregu rozdzielczego nie jest jednoznaczne.**
2. **Pomiędzy klasami nie może być luk (każdy element szeregu musi należeć do pewnej klasy. Szczególną uwagę należy zwrócić na pierwszą i ostatnią klasę (należy ewentualnie je powiększyć). Liczba elementów w każdej klasie nie może być zbyt mała, ani zbyt duża.**
3. **Zastępując szereg statystyczny szeregiem rozdzielczym tracimy dokładność pewnych informacji. W pewnych praktycznych zastosowaniach nie jest to istotne.**
4. **Liczba klas zazwyczaj nie przekracza 30.**

<u>Średnia arytmetyczna</u>

i. **Szereg rozdzielczy**

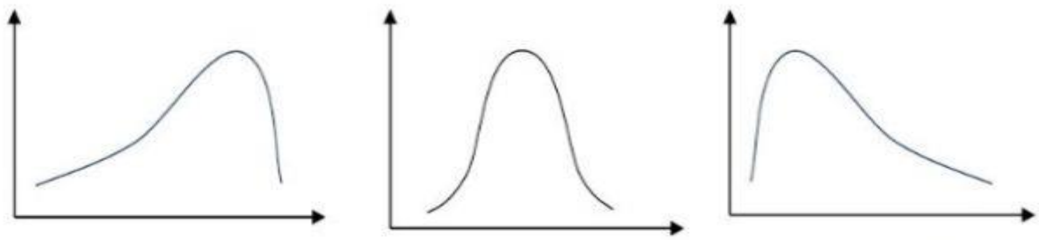
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i^s, \qquad x_i^s - \text{środek klasy}$$

<u>Wariancja</u> *(oznaczana przez s^2, σ^2, D^2)*

A) **Szereg rozdzielczy**

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i^s - \overline{x})^2 * n_i, \qquad x_i^s - \text{środek klasy}$$

Asymetria



Rozkład asymetryczny, lewostronny, ujemny

Rozkład symetryczny

Rozkład asymetryczny, prawostronny, dodatni

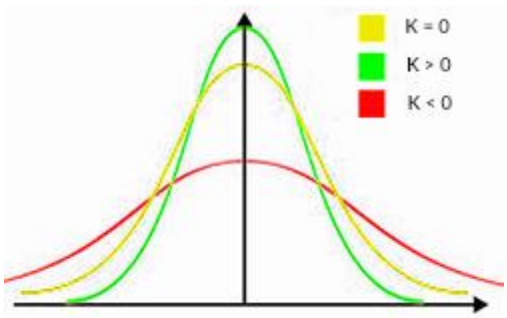
$\boldsymbol{\bar{x} < M < D}$ $\boldsymbol{\bar{x} > M > D}$ $\boldsymbol{\bar{x} = M = D}$

Współczynnik koncentracji (skupienia), kurtoza K

$$K = \frac{M_4}{s^4} - 3$$

gdzie $\boldsymbol{M_l = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^l}$**, a s jest odchyleniem standardowym.**

Kurtoza ma wartość 0 dla rozkładu normalnego. Gdy K>0, to rozkład jest bardziej skupiony („szpiczasty”) niż rozkład normalny, gdy K<0, to rozkład jest bardziej spłaszczony niż rozkład normalny.



$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ standardowy rozkład normalny