Marek Cieciura

METODY PROBABILISTYCZNE

WYBRANE TWIERDZENIA

Twierdzenia matematyczne uważane są za prawdziwe, albowiem w niczyim interesie nie leży, by uważać je za fałszywe - Monteskiusz (1689-1755), francuski myśliciel, prawnik, pisarz

Warszawa, marzec 2022

SPIS TREŚCI

Spis treści

UWAGI WSTĘPNE[1]

Statystyka dostarcza metod, za pomocą których można weryfikować dane empiryczne, tworzyć modele matematyczne zjawisk oraz sporządzać rozmaite prognozy.

Nauczanie statystyki powinno stanowić coś więcej, niż dostarczanie rutynowych rozwiązań. Powinno rozszerzyć analityczne umiejętności studentów oraz zwiększyć poziom zdrowego sceptycyzmu i rozwinąć innowacyjność.

W szczególności statystyka stanowi w nauczaniu studentów nie tylko źródło praktycznej wiedzy, ale powinna spowodować krytyczne podejście na analizowane problemy. Wywołanie nawyku dowodzenia twierdzeń uważanych za intuicyjnie oczywiste i ukazanie, że nie wszystko co pozornie oczywiste jest prawdziwe stanowi być może najważniejszy rezultat studiowania statystyki.

Można spotkać pogląd, by na studiach informatycznych koncentrować się na metodach i wzorach, a dowody pomijać. Pomijając dowody studenci przyzwyczajają się do tego, by bezkrytycznie akceptować czyjeś stwierdzenia i czuć się zwolnionym z potrzeby poszukiwania prawdy.

Postęp w dostępności programów komputerowych drastycznie zmienił sposób stosowania metod statystycznych. Jeszcze dwadzieścia lat temu studenci rozwiązywali zadania z programowania liniowego używając metody simpleks z pomocą ręcznego kalkulatora. Dziś rzadko wymaga się szczegółowej znajomości metod, nie mówiąc o dowodach ich poprawności. Wszyscy pozostali uczą się stosowania modułów analizy statystycznej oferowanych przez wiele programów. Przykładowo szacowanie współczynników regresji nie wymaga więcej wysiłku niż wybranie ikony na ekranie komputera.

Dostępność programów wykonujących skomplikowane obliczenia natychmiast i to bezbłędnie ogromnie przyspieszyła cały proces analizy. W konsekwencji modele, które dawniej były zbyt skomplikowane by je badać, mogą być dziś rozwiązywane przez badaczy obeznanych z komputerem.

Komputery nie zmieniły jednak roli jako narzędzia do rozwinięcia argumentu i wykroczenia poza to, co wydaje się zgodne ze zdrowym rozsądkiem i/lub polityczną poprawnością, nie zmieniła się. Dostępność oprogramowania pomagającego badaczom stosować wzory i procedury statystyczne nie umniejsza roli statystyki jako treningu umysłowego.

Nie ulega wątpliwości, że studenci informatyki powinni uczyć się statystyki. Spełnia to podwójną rolę:

- docenienie siły umysłu, informatycy pozbawieni takiego treningu mogą łatwiej akceptować półprawdy i przeoczać braki w logice swoich argumentów poprzez nauczanie statystyki w odpowiednim tempie, nie omijając dowodów i przywiązując właściwą uwagę do spójności całej teorii.
- poznanie konkretnych technik pomagających zauważać wzajemne powiązania i wyciągać wnioski poprzez zwracanie uwagi, jakie problemy mogą rozwiązać przy pomocy danej techniki; wymaga też, by przywiązywać dostateczną uwagę do formułowania pytań używając formalnego języka. Zarówno w przeszłości, jak i teraz statystyka bywa niewłaściwie wykorzystywana przez stosowanie wyrafinowanych technik do kiepsko zidentyfikowanych problemów.

[1] Opracowano na podstawie: Tomasz Żylicz Czy należy uczyć matematyki studentów ekonomii?, lipiec 2007, http://coin.wne.uw.edu.pl/tzylicz/matekon.pdf

1. STATYSTYKA OPISOWA

1.1. Własność 1 średniej arytmetycznej elementów próby

Wykazać, że średnia arytmetyczna elementów próby $(x_1, x_2, ..., x_n)$ spełnia zależność:

$$x_{min} \le \overline{x}_n \le x_{max} \tag{1-1.1}$$

1.2. Własność 2 średniej arytmetycznej elementów próby

Wykazać, że średnia arytmetyczna elementów próby $(x_1, x_2, ..., x_n)$ spełnia zależność:

$$\sum_{i=1}^{n} (x_i - \overline{x}_n) = 0 \tag{1-2.1}$$

1.3. Własność 3 średniej arytmetycznej elementów próby

Wykazać, że średnia arytmetyczna elementów próby $(x_1, x_2, ..., x_n)$ spełnia zależność

$$\sum_{x_i < \overline{x}_n} (\overline{x}_n - x_i) = \sum_{x_i > \overline{x}_n} (x_i - \overline{x}_n) \tag{1-3.1}$$

1.4. Własność 4 średniej arytmetycznej elementów próby

Wykazać, że dla elementów próby $(x_1, x_2, ..., x_n)$ wyrażenie

$$\sum_{i=1}^{n} (x_i - c)^2 \tag{1-4.1}$$

ma wartość najmniejszą gdy c= $\overline{x}_n$

1.5. Relacje pomiędzy średnimi

Wykazać prawdziwość zależności:

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \le \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \le \frac{1}{n}\sum_{i=1}^{n} x_i \le \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{1-5.1}$$

dla n=2

1.6. Wyznaczanie wariancji z próby

Wykazać, że wariancję z próby można wyznaczyć ze wzoru

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 \tag{1-6.1}$$

2. RACHUNEK PRAWDOPODOBIEŃSTWA

2.1. Dopełnienie sumy zdarzeń – 1 prawo de Morgana

Wykazać, że dopełnienie sumy zdarzeń jest równe iloczynowi ich dopełnień

$$(A \cup B)' = A' \cap B' \tag{2-1.1}$$

2.2. Dopełnienie iloczynu zdarzeń – 2 prawo de Morgana

Wykazać, że dopełnienie iloczynu dwóch zdarzeń jest równe sumie ich dopełnień

$$(A \cap B)' = A' \cup B' \tag{2-2.1}$$

2.3. Prawdopodobieństwo zdarzenia niemożliwego

Wykazać, że prawdopodobieństwo zdarzenia niemożliwego jest równe zeru

$$P(\emptyset) = 0 \tag{2-3.1}$$

2.4. Monotoniczność prawdopodobieństwa

Wykazać, że jeżeli zdarzenie A pociąga zdarzenie B, to prawdopodobieństwo zdarzenia A jest nie większe niż prawdopodobieństwo zdarzenia B

$$A \subset B \Rightarrow P(A) \le P(B) \tag{2-4.1}$$

2.5. Prawdopodobieństwo zdarzenia przeciwnego

Wykazać, że prawdopodobieństwa zdarzenia przeciwnego A' jest równe dopełnieniu do 1 prawdopodobieństwa zdarzenia A

$$P(A') = 1 - P(A) \tag{2-5.1}$$

2.6. Prawdopodobieństwo sumy dwóch zdarzeń

Wykazać, że prawdopodobieństwo sumy dwóch zdarzeń jest równe sumie prawdopodobieństw tych zdarzeń zmniejszonej o prawdopodobieństwo ich iloczynu

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2-6.1}$$

2.7. Niezależność zdarzeń przeciwnych

Wykazać, że jeżeli zdarzenia A i B są niezależne, to A' i B są również niezależne

$$P(A \cap B) = P(A) \cdot P(B) \Rightarrow P(A' \cap B) = P(A') \cdot P(B) \tag{2-7.1}$$

2.8. Prawdopodobieństwo całkowite

Jeżeli zdarzenia $A_1, A_2, \dots, A_k$ o dodatnich prawdopodobieństwach wykluczają się parami i suma ich jest zdarzeniem pewnym, to dla dowolnego zdarzenia B zachodzi wzór

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_k)P(B/A_k) = \sum_{i=1}^{k} P(B/A_i) \cdot P(A_i) \tag{2-8.1}$$

2.9. Wzór Bayesa

Wykazać, że jeżeli zdarzenia $A_1, A_2, \dots, A_k$ o dodatnich prawdopodobieństwach wykluczają się parami i suma ich jest zdarzeniem pewnym, zaś B jest dowolnym zdarzeniem o dodatnim prawdopodobieństwie, to zachodzi wzór

$$P(A_j / B) = \frac{P(A_i) \cdot P(B / A_i)}{\sum_{i=1}^{k} P(B / A_i) \cdot P(A_i)} \text{ dla j=1, ... , k} \quad (2\text{-}9.1)$$

2.10. Wartość oczekiwana iloczynu stałej i zmiennej losowej – jednorodność

Wykazać, że

$$E(aX) = aEX \quad (2\text{-}10.1)$$

gdzie a – stała

2.11. Wartość oczekiwana sumy zmiennych losowych - addytywność

Wykazać, że jeśli X i Y są zmiennymi losowymi o łącznej funkcji gęstości f(x, y) - lub w przypadku zmiennych skokowych o łącznej funkcji prawdopodobieństwa $p(x_i, y_j)$-wówczas wartość oczekiwana sumy Z = X + Y jest równa sumie wartości oczekiwanych (jeśli one istnieją) zmiennych X i Y

$$E(X + Y) = EX + EY \quad (2\text{-}11.1)$$

2.12. Wartość oczekiwana iloczynu zmiennych losowych

Wykazać, że jeżeli dwie zmienne losowe X i Y są niezależne to wartość oczekiwana ich iloczynu jest równa iloczynowi ich wartości oczekiwanych, co zapisujemy

$$E(X \cdot Y) = EX \cdot EY \quad (2\text{-}12.1)$$

2.13. Wariancja sumy niezależnych zmiennych losowych

Wykazać, że jeżeli zmienne losowe są niezależne, to wariancja ich sumy jest równa sumie ich wariancji

$$D^2(X+Y) = D^2X + D^2Y \quad (2\text{-}13.1)$$

2.14. Parametry rozkładu standaryzowanej zmiennej losowej

Wykazać, że jeżeli zmienna losowa X ma wartość oczekiwaną EX=m i odchylenie standardowe DX=$\sigma > 0$, to zmienna standaryzowana Y

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

ma wartość oczekiwaną równą zeru i odchylenie standardowe równe jeden

$$EY=0, \quad DY=1$$

$$X : N(m, \sigma) \Rightarrow Y = \frac{X - m}{\sigma} : N(0,1) \quad (2\text{-}14.1)$$

2.15. Rozkład sumy niezależnych zmiennych losowych o rozkładach normalnych

Wykazać, że jeżeli X i Y są niezależnymi zmiennymi losowymi, przy czym X: $N(m_x, \sigma_x)$ i Y: $N(m_y, \sigma_y)$ to ich suma Z = X + Y ma rozkład $N(m_x+m_y, \sqrt{\sigma_x^2 + \sigma_y^2})$.

$$X : N(m_x, \sigma_x)\ Y : N(m_y, \sigma_y), f_3(x, y) = f_1(x) \cdot f_2(y) \Rightarrow X + Y : N(m_x + m_y, \sqrt{\sigma_x^2 + \sigma_y^2}) \quad (2\text{-}15.1)$$

2.16. Własności współczynnika korelacji

1) Współczynnik korelacji spełnia podwójną nierówność:

$$-1 \le \rho \le 1 \quad (2\text{-}16.1)$$

2) Warunkiem koniecznym i wystarczającym na to, aby

$$P(Y = a \cdot X + b) = 1 \text{ gdzie } a \ne 0 \quad (2\text{-}16.2)$$

jest

$$\rho^2 = 1 \qquad (2\text{-}16.3)$$

co można zapisać w postaci

$$\rho^2 = 1 \Leftrightarrow P(Y = a \cdot X + b) = 1 \text{ gdzie } a \neq 0 \qquad (2\text{-}16.4)$$

2.17. Funkcja regresji drugiego rodzaju

Wykazać, że funkcja $Y = a \cdot X + b$ jest funkcją regresji 2 rodzaju, tzn. zapewnia minimum wyrażenia

$$E[Y - (aX - b)]^2 \to \min \qquad (2\text{-}17.1)$$

jeżeli jej współczynniki są równe

$$\hat{a} = \rho \frac{\sigma_x}{\sigma_y} \text{ oraz } \hat{b} = m_x - \rho \frac{\sigma_x}{\sigma_y} m_y$$

tzn. ma postać

$$Y = \rho \frac{\sigma_x}{\sigma_y} \cdot X + m_x - \rho \frac{\sigma_x}{\sigma_y} m_y \qquad (2\text{-}17.2)$$

2.18. Twierdzenie Lindeberga-Levy'ego

Rozpatrzmy ciąg zmiennych losowych $\{X_n\}$ o których zakładamy, ze:
1) są to zmienne niezależne
2) mają identyczne rozkłady
3) mają skończone wartości oczekiwane m i skończone wariancje σ^2

Wprowadźmy ciąg zmiennych losowych $\{Y_n\}$, gdzie zmienna losowa Y_n jest sumą

Na podstawie **(**2-12.1)

$$EY_n = E\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} m = n \cdot m$$

a na podstawie (2-13.1)

$$D^2Y_n = D^2\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} D^2X_i = \sum_{i=1}^{n} \sigma^2 = n \cdot \sigma^2$$

W końcu dokonajmy standaryzacji zmiennej losowej Y_n

$$V_n = \frac{Y_n - n \cdot m}{\sigma\sqrt{n}}$$

i oznaczmy przez $\{F_n(v)\}$ ciąg dystrybuant ciągu zmiennych losowych V_n.

Twierdzenie Lindeberga-Levy'ego orzeka, ze przy powyższych założeniach mamy

$$\lim_{n\to\infty} F_n(v) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} e^{-\frac{z^2}{2}} dz$$

3. STATYSTYKA MATEMATYCZNA

3.1. Rozkład średniej arytmetycznej elementów próby

Wykazać, że średnia arytmetyczna elementów próby $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ gdy X_i: N(m,σ) ma rozkład

$\overline{X}_n : N(m, \frac{\sigma}{\sqrt{n}})$

$$X_i : N(m,\sigma) \Rightarrow \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i : N(m, \frac{\sigma}{\sqrt{n}}) \tag{3-1.1}$$

3.2. Estymator nieobciążony wartości oczekiwanej

Wykazać, że średnia arytmetyczna z próby jest estymatorem nieobciążonym wartości oczekiwanej cechy populacji.

$$E\overline{X}_n = m \tag{3-2.1}$$

3.3. Estymator najefektywniejszy wartości oczekiwanej

Wykazać, że średnia arytmetyczna z próby $\overline{X}_n$ jest estymatorem najefektywniejszym wartości oczekiwanej m cechy populacji o rozkładzie N(m,σ).

$$X : N(m,\sigma) \Rightarrow D^2\overline{X}_n \to \min \tag{3-3.1}$$

3.4. Obciążoność wariancji z próby

Wykazać, że wariancja z próby S_n^2 jest estymatorem obciążonym wariancji σ^2 cechy populacji.

$$ES_n^2 = \sigma^2 \tag{3-4.1}$$

3.5. Estymator wartości oczekiwanej rozkładu normalnego – metoda największej wiarogodności

Wykazać, że metodą największej wiarogodności otrzymuje się estymator wartości oczekiwanej cechy X o rozkładzie N(m,σ) w postaci

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3-5.1}$$

3.6. Przedział ufności dla wartości oczekiwanej rozkładu normalnego przy znanej wariancji

Wykazać, że przedział ufności dla wartości oczekiwanej rozkładu normalnego ma postać

$$\left[\overline{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{3-6.1}$$

gdzie: u_α wyznacza się na podstawie dystrybuanty rozkładu normalnego Φ: $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$

3.7. Weryfikacja hipotezy o równości wartości oczekiwanych cech o rozkładach normalnych przy znanych odchyleniach standardowych

Wykazać, że w zadaniu weryfikacji hipotezy o równości wartości oczekiwanych cech o rozkładach normalnych przy znanych odchyleniach standardowych hipotezę odrzuca się przy spełnieniu nierówności $\left|\bar{X}_{n_1} - \bar{Y}_{n_2}\right| > k_\alpha$, gdzie k_α wyznacza się w odpowiedni sposób z tablic rozkładu normalnego.

3.8. Współczynnik korelacji Spearmana

Uzasadnić postać wzoru na współczynnik korelacji Spearmana

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(c_i - d_i)^2}{n(n^2-1)} \tag{3-8.1}$$

i dokonać jego analizy.

3.9. Współczynnik korelacji punktowo-dwuseryjnej

Uzasadnić postać wzoru na współczynnik korelacji punktowo-dwuseryjnej

$$r_{xy} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\sqrt{p^\circ q}$$

i dokonać jego analizy.

3.10. Wyznaczanie współczynników regresji metodą najmniejszych kwadratów

Wykazać, że metodą najmniejszych kwadratów uzyskuje się następujące współczynniki w równaniu regresji jednej zmiennej $y = a \cdot x + b$

$$\hat{a} = r\frac{s_y}{s_x} \quad \hat{b} = \bar{y}_n - r\frac{s_y}{s_x}\bar{x}_n \tag{3-10.1}$$

3.11. Wyznaczanie współczynników regresji metodą największej wiarogodności

Wykazać, że metodą największej wiarogodności gdy dla każdego x cecha Y ma rozkład normalny N(ax+b,σ), tzn. gęstość zmiennej losowej przy ustalonej wartości x ma postać

$$f_x(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y-ax-b)^2}{2\sigma^2}}$$

uzyskuje się następujące współczynniki w równaniu regresji jednej zmiennej $y = a \cdot x + b$

$$\hat{a} = r\frac{s_y}{s_x} \quad \hat{b} = \bar{y}_n - r\frac{s_y}{s_x}\bar{x}_n \tag{3-11.1}$$