

Prawdopodobieństwo warunkowe

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Wzór Bayesa (twierdzenie o prawdopodobieństwie całkowitym)

Jeśli B' to zdarzenie dopełniające dla B , to:

$$P(A) = P(B)P(A|B) + P(B')P(A|B')$$

Przykład 1

Na pewną chorobę choruje 30% populacji. Test do wykrywania choroby ma czułość 99.7% i swoistość 98.5%. Jakie jest prawdopodobieństwo, że osoba z wynikiem pozytywnym faktycznie jest chora?

- czułość testu (sensitivity) – prawdopodobieństwo pozytywnego wyniku u osoby chorej;
- swoistość testu (specificity) – prawdopodobieństwo negatywnego wyniku u osoby zdrowej;
- wynik *falszywie pozytywny* – pozytywny wynik u osoby zdrowej;
- wynik *falszywie negatywny* – negatywny wynik u osoby chorej;

A – osoba chora, B – wynik pozytywny testu

$$P(A) = 0.3$$

$$P(B|A) = 0.997 \text{ oraz } P(B'|A) = 0.003$$

$$P(B'|A') = 0.985 \text{ oraz } P(B|A') = 0.015$$

$$P(A|B) = ?$$

$$P(A \cap B) = P(B|A) \cdot P(A) = 0.997 \cdot 0.3 = 0.2991$$

$$P(B) = P(A)P(B|A) + P(A')P(B|A') = 0.3 \cdot 0.997 + 0.7 \cdot 0.015 = 0.3096$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2991}{0.3096} = 96.6\%$$

Co jeśli chore jest 1% populacji?

Wówczas $P(A \cap B) = 0.00997$, $P(B) = 0,02482$, $P(A|B) = 40.2\%$

Próba Bernoulliego

Pewien eksperyment ma dwa możliwe wyniki: sukces lub porażka. Prawdopodobieństwo sukcesu w każdej próbie wynosi p . Jeśli Y to zmienna losowa oznaczająca liczbę sukcesów w n (niezależnych) próbach, to:

$$P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$EY = EY_1 + \dots + EY_n = p + \dots + p = np$$

$$\text{Var}(Y) = \text{Var}(Y_1) + \dots + \text{Var}(Y_n) = p(1-p) + \dots + p(1-p) = np(1-p)$$

Przykład 2

Rzucamy 5 razy monetą i liczymy liczbę otrzymanych orłów (zmienna Y). Obliczyć prawdopodobieństwa wyrzucenia dokładnie 2 orłów oraz co najmniej 4 orłów.

$$P(Y = 2) = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{5}{16}$$

$$P(Y = 4) = \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 = \frac{5}{32}$$

$$P(Y = 5) = \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

$$P(Y \geq 4) = P(Y = 4) + P(Y = 5) = \frac{3}{16}$$

Przykład 3

Jeden na ośmiu dorosłych mężczyzn ma podniesiony poziom cholesterolu. Losowo wybrano 6 mężczyzn z populacji. Jakie jest prawdopodobieństwo, że dokładnie 2 spośród nich ma podniesiony poziom cholesterolu? Jakie jest prawdopodobieństwo, że co najmniej jeden z nich ma podniesiony poziom cholesterolu?

$$P(Y = 2) = \binom{6}{2} \left(\frac{1}{8}\right)^2 \left(\frac{7}{8}\right)^4 = 13.7\%$$

$$P(Y = 0) = \binom{6}{0} \left(\frac{1}{8}\right)^0 \left(\frac{7}{8}\right)^6 = 44.9\%$$

$$P(Y \geq 1) = 1 - P(Y = 0) = 55.1\%$$

Przybliżanie rozkładu Bernoulliego rozkładem normalnym

Rozkład Bernoulliego z parametrami n i p przybliżamy rozkładem normalnym z $\mu = np$ i $\sigma = \sqrt{np(1-p)}$. Otrzymane przybliżenie:

- ułatwia rachunki (wyliczanie współczynników Newtona jest trudne dla dużych n i k);
- jest dość dobre, gdy $np \geq 5$ i $n(1-p) \geq 5$, przy czym przybliżenie jest tym lepsze, im p jest bliższe 0.5;
- jest dość dokładne w centrum rozkładu i gorsze w ogonach;
- jest przybliżeniem zmiennej dyskretnej przy pomocy zmiennej ciągłej, więc przyjmujemy, że $Y = k$ odpowiada $k - 0.5 \leq Y' \leq k + 0.5$

Przykład 4

Zmienna losowa Y ma rozkład Bernoulliego z parametrami $n = 40$ i $p = 0.25$. Chcemy policzyć prawdopodobieństwo $P(10 \leq Y \leq 15)$

Przybliżamy zmienną Y' o rozkładzie normalnym o parametrach $\mu = np = 10$ i $\sigma = \sqrt{np(1-p)} = 2.74$. Normalizujemy:

$Z = \frac{Y' - 10}{2.74}$ i obliczamy:

$$P(10 \leq Y \leq 15) = P(9.5 \leq Y' \leq 15.5)$$

$$\begin{aligned} P\left(\frac{9.5-10}{2.74} \leq Z \leq \frac{15.5-10}{2.74}\right) &= P(-0.18 \leq Z \leq 2.01) \\ &= 0.9778 - 0.4286 = 0.539 \text{ (prawdziwa wartość to } 0.534) \end{aligned}$$

Przykład 5

Wynikiem testu jest liczba całkowita z przedziału $[0, 150]$. Wyniki testu mają rozkład normalny z parametrami $\mu = 100$ i $\sigma = 16$. Wyznaczyć prawdopodobieństwo, że losowo wybrany student uzyskał wynik między 120 a 140 punktów.

$$\begin{aligned} P(120 \leq Y \leq 140) &= P(119.5 \leq Y' \leq 140.5) \\ P\left(\frac{119.5-100}{16} \leq Z \leq \frac{140.5-100}{16}\right) &= P(1.22 \leq Z \leq 2.53) \\ &= 0.9943 - 0.8888 = 0.1055 \end{aligned}$$

Rozkłady próbkowe

Rozważmy populację o pewnym rozkładzie, np.

- normalnym $N(\mu, \sigma)$ z parametrami μ i σ
- dwupunktowym z parametrem p
(tzn. $P(Y = 1) = p$, $P(Y = 0) = 1 - p$)

Wybieramy próbę o rozmiarze n z populacji otrzymując:

- wartości $y_1, \dots, y_n$
- łączną liczbę sukcesów y

Obliczamy estymatory:

- $\bar{y} = \frac{y_1 + \dots + y_n}{n}$ oraz $s = \sqrt{\frac{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n-1}}$
- $\hat{p} = \frac{y}{n}$

Gdy n jest duże, estymatory są na ogół bliskie parametrom.

- Jak bardzo estymatory mogą różnić się od prawdziwych parametrów?
- Co się stanie, gdy wylosujemy inną próbę?

Wyobraźmy sobie, że losujemy próbę wiele razy. Interesuje nas rozkład wszystkich możliwych do uzyskania wartości $\bar{y}$, s lub $\hat{p}$. Taki rozkład nazywamy *rozkładem próbkowym* estymatora.

Przykład 6

Producent ocenia, że 2% jego wyrobów jest wadliwych. Wyroby pakowane są po 40 sztuk w jednym opakowaniu.

Y to liczba wadliwych rozkładów w losowo wybranej paczce ma rozkład dwumianowy (rozkład Bernoulliego) z parametrami $n = 40$, $p = 0.02$).

Niech $\hat{p} = Y/40$ to frakcja elementów wadliwych.

$$P(\hat{p} = 0) = P(Y = 0) = 1 \cdot (0.02)^0 \cdot (0.98)^{40} = 0.45$$

$$P(\hat{p} = 0.025) = P(Y = 1) = 40 \cdot (0.02)^1 \cdot (0.98)^{39} = 0.36$$

$$P(\hat{p} = 0.050) = P(Y = 2) = 780 \cdot (0.02)^2 \cdot (0.98)^{38} = 0.14$$

$$P(\hat{p} = 0.075) = P(Y = 3) = 9880 \cdot (0.02)^3 \cdot (0.98)^{37} = 0.04$$

$$P(\hat{p} \geq 0.1) = P(Y \geq 4) = 0.01$$

Estymator $\hat{p}$

Niech Y ma rozkład Bernoulliego (n, p) . Wówczas:

$$\mu_Y = np \text{ oraz } \text{Var}Y = np(1 - p)$$

Dla estymatora $\hat{p} = Y/n$ mamy:

$$\mu_{\hat{p}} = \frac{np}{n} = p \text{ oraz } \text{Var}\hat{p} = \frac{np(1 - p)}{n^2} = \frac{1}{n}p(1 - p)$$

np. dla $p = 0.3$ otrzymujemy:

$$P(0.25 \leq \hat{p} \leq 0.35) = 0.500 \quad \text{gdy } n = 10$$

$$P(0.25 \leq \hat{p} \leq 0.35) = 0.535 \quad \text{gdy } n = 20$$

$$P(0.25 \leq \hat{p} \leq 0.35) = 0.612 \quad \text{gdy } n = 40$$

$$P(0.25 \leq \hat{p} \leq 0.35) = 0.728 \quad \text{gdy } n = 80$$

$$P(0.25 \leq \hat{p} \leq 0.35) = 0.987 \quad \text{gdy } n = 500$$

Dana jest próba o rozmiarze n z populacji o rozkładzie normalnym. Obserwujemy średnią próbkową $\bar{y}$. Jaki jest rozkład tych średnich przy wielokrotnym próbkowaniu?

Fakt 1

Suma zmiennych niezależnych o rozkładzie normalnym ma rozkład normalny.

Fakt 2

Jeśli X ma rozkład normalny, to $aX + b$ dla $a \neq 0$ też ma rozkład normalny.

Wniosek

Jeśli Y ma rozkład normalny $N(\mu, \sigma)$, to rozkład próbkowy $\bar{Y}$ jest normalny $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Centralne Twierdzenie Graniczne

Centralne Twierdzenie Graniczne

Jeśli $Y_1, \dots, Y_n$ to niezależne zmienne losowe o jednakowym rozkładzie i skończonej wariancji, to dla dużego n zmienna losowa $\bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n)$ ma rozkład bliski normalnemu.

- jeśli rozkład Y jest normalny, to dla dowolnego n rozkład $\bar{Y}$ jest normalny;
- jeśli rozkład Y jest w przybliżeniu symetryczny i nie ma „ciężkich ogonów”, to $n = 30$ jest wystarczająco duże.