Wykład z analizy danych: wyprowadzenie prostego testu w rozkładzie dwupunktowym

Marek Kubiak

Instytut Informatyki
Politechnika Poznańska

Model statystyczny

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą losową prostą z rozkładu zero–jedynkowego $B_1(p)$. Wartość p jest nieznana.

Interesuje nas weryfikacja hipotezy statystycznej:

$$H_0 : p = p_0$$

Inna możliwość, która nas interesuje (alternatywa), to:

$$H_1 : p < p_0$$

Weryfikacja ta ma być przeprowadzona przy prawdopodobieństwie uzyskania błędnych wniosków ograniczonym do α (np. 0,05; 0,01).

Przykład banku

Z dotychczasowej historii kilkuletniego stosowania pewnej procedury klasyfikacji kredytobiorców w banku wynika, że 9% udzielonych kredytów jest w ogóle niespłacanych.

Bank rozważa decyzję wprowadzenia do użytku nowej, udoskonalonej procedury klasyfikowania wniosków kredytowych.

Na podstawie testów nowej procedury na grupie 130 losowo wybranych, nowych klientów należy stwierdzić, czy należy wprowadzić nową procedurę do powszechnego użycia.

Przykład banku

Dysponujemy próbą losową prostą $\mathbf{X} = (X_1, \ldots, X_{130})$ (czyli $n = 130$) z rozkładu $B_1(p)$.

Interesuje nas weryfikacja hipotezy statystycznej:

$$H_0 : p = p_0 = 0,09$$

Inna możliwość, która nas interesuje (alternatywa), to:

$$H_1 : p < p_0 = 0,09$$

Weryfikacja ta ma być przeprowadzona przy prawdopodobieństwie uzyskania błędnych wniosków ograniczonym do $\alpha = 0,05$.

Statystyka testowa i jej rozkład

Wiemy (tw. Bernoulliego), że statystyka:

$$S_n = \sum_{i=1}^{n} X_i \ \sim \ B_n(p)$$

Możemy obliczyć:

$$E(S_n) = n \cdot p$$

$$D^2(S_n) = n \cdot p \cdot (1-p)$$

$$D(S_n) = \sqrt{n \cdot p \cdot (1-p)}$$

Wstępne przyjęcie prawdziwości H_0

Wstępnie przyjmujemy: $H_0 : p = p_0$

Tylko dzięki temu założeniu **ustalony jest rozkład** statystyki S_n, możemy obliczyć wartości $E(S_n)$ i $D(S_n)$ i prowadzić wnioskowanie dalej!

Faktycznie dążymy do zaprzeczenia H_0 – przyjmujemy ją wstępnie po to, żeby pokazać, że doprowadza do nieprawdopodobnych wniosków!

Przykład banku (c.d.)

Wstępnie przyjmujemy: $H_0 : p = p_0 = 0,09$

Obliczamy:

$$E(S_{130}) = n \cdot p_0 = 130 \cdot 0,09 = 11,7$$

$$D(S_{130}) = \sqrt{n \cdot p_0 \cdot (1 - p_0)} = \sqrt{130 \cdot 0,09 \cdot (1 - 0,09)} \approx 3,26$$

Statystyka testowa – standaryzacja

Ustandaryzujmy statystykę S_n:

$$Z = \frac{S_n - E(S_n)}{D(S_n)} = \frac{S_n - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$$

Zgodnie z tw. Lindeberga–Levy'ego
(przy $n \cdot p \geq 5$ i $n \cdot (1 - p) \geq 5$),
statystyka Z ma w granicy standaryzowany rozkład normalny:

$$Z \sim N(0, 1)$$

Przykład banku (c.d.)

Sprawdzamy spełnienie warunków dobrego przybliżenia:

$$n \cdot p = 130 \cdot 0,09 = 11,7 \geq 5$$

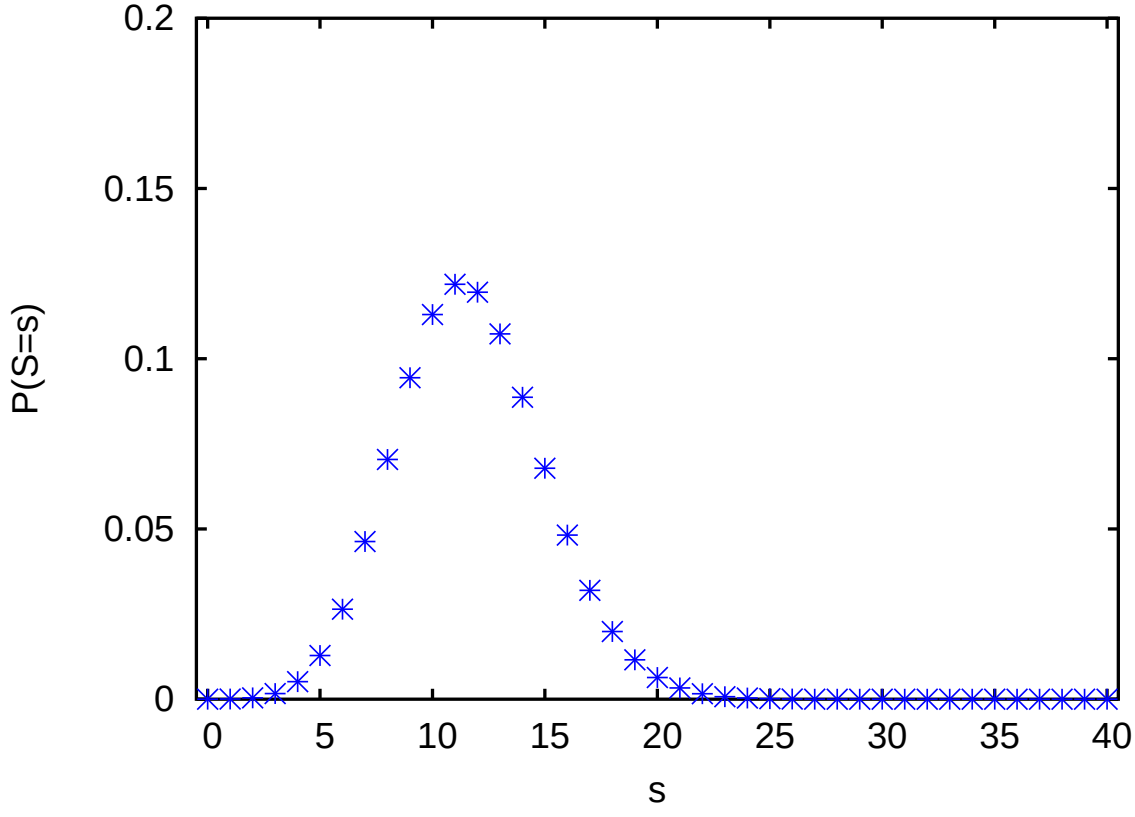
$$n \cdot (1-p) = 130 \cdot 0,91 = 118,3 \geq 5$$

W takim razie możemy z wystarczająco dużą dokładnością przyjąć:

$$Z = \frac{S_{130} - E(S_{130})}{D(S_{130})} = \frac{S_{130} - 11,7}{3,26} \sim N(0,1)$$

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

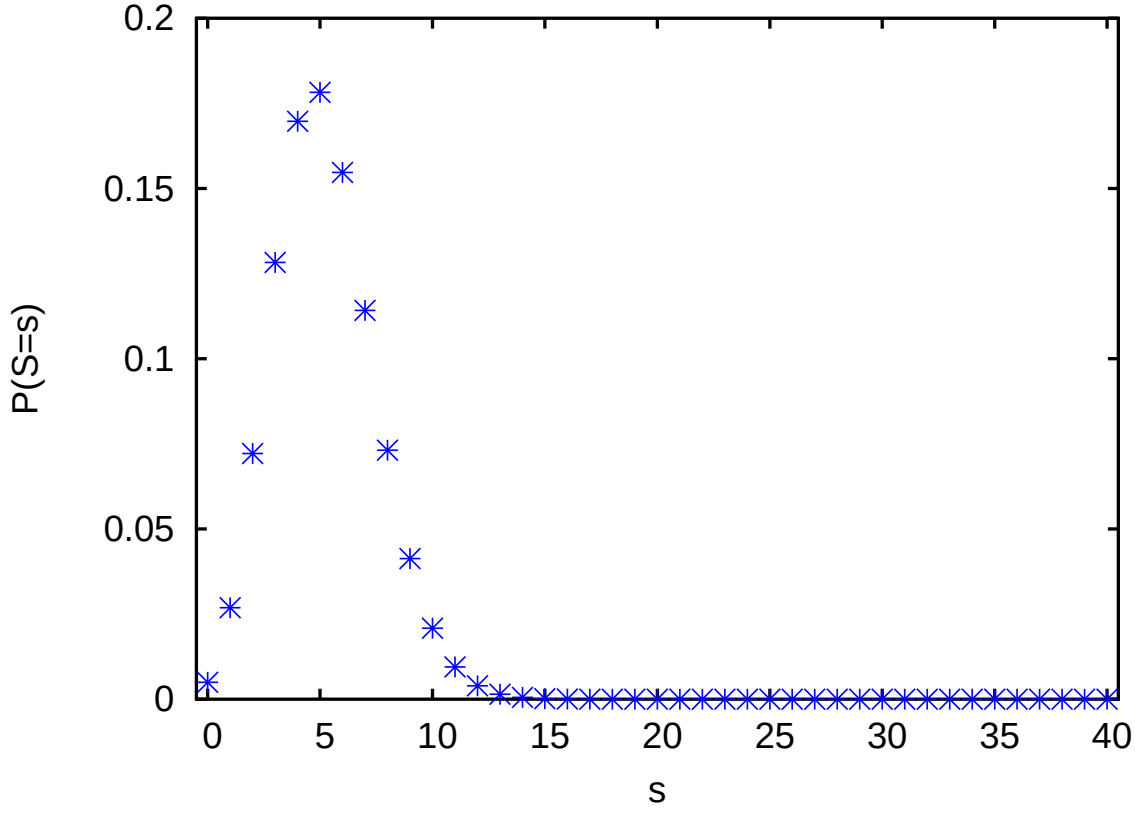
Co możemy powiedzieć o hipotezach H_0 i H_1 po obliczeniu wartości statystyki S_n na podstawie pobranej próbki?



Rysunek: Rozkład prawdopodobieństwa $B_{130}(0,09)$

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

Co możemy powiedzieć o hipotezach H_0 i H_1 po obliczeniu wartości statystyki S_n na podstawie pobranej próbki?



Rysunek: Rozkład prawdopodobieństwa $B_{130}(0,04)$

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

Co możemy powiedzieć o hipotezach H_0 i H_1 po obliczeniu wartości statystyki S_n na podstawie pobranej próbki?

$$S_n \gg E(S_n) \text{ przeczy } H_0 \text{ i przeczy } H_1$$

$$S_n \approx E(S_n) \text{ sprzyja } H_0 \text{ i przeczy } H_1$$

$$S_n \ll E(S_n) \text{ przeczy } H_0 \text{ i sprzyja } H_1$$

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

Co z tego wynika dla statystyki:

$$Z = \frac{S_n - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$$

$$Z \gg 0 \text{ przeczy } H_0 \text{ i przeczy } H_1$$

$$Z \approx 0 \text{ sprzyja } H_0 \text{ i przeczy } H_1$$

$$Z \ll 0 \text{ przeczy } H_0 \text{ i sprzyja } H_1$$

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

Dążymy do zaprzeczenia H_0 i przyjęcia H_1, więc **zbiorem krytycznym** (zbiorem odrzuceń) dla hipotezy H_0 (z alternatywą H_1) będzie:

$$Z < z_\alpha < 0$$

dla którego:

$$P(Z < z_\alpha) = \alpha$$

Jest to zbiór wartości statystyki Z, który sprzyja H_1, a przy prawdziwości H_0 jest *mało prawdopodobny*!

Określenie zbioru krytycznego statystyki testowej (H_0 prawdziwa)

Szukamy *wartości krytycznej* z_α (ujemnej):

$$P(Z < z_\alpha) = \alpha$$

$$\Phi(z_\alpha) = \alpha$$

$$1 - \Phi(-z_\alpha) = \alpha$$

$$\Phi(-z_\alpha) = 1 - \alpha$$

Przykład banku (c.d.)

Obliczenie wartości krytycznej z_α:

$$P(Z < z_\alpha) = \alpha = 0,05$$

$$\Phi(-z_\alpha) = 1 - \alpha = 0,95$$

(odczytujemy z tablic)

$$-z_\alpha = 1,65$$

$$z_\alpha = -1,65$$

Zbiór krytyczny – podsumowanie

Obliczamy wartość z statystyki testowej Z na podstawie próbki i porównujemy obliczoną wartość z wartością krytyczną:

- jeśli $z < z_\alpha$, to **odrzucamy** H_0 i **przyjmujemy** H_1 (zaszło zdarzenie mało prawdopodobne dla założonej H_0)
- jeśli $z \geq z_\alpha$, to **brak podstaw do odrzucenia** H_0

Test statystyczny – podsumowanie

Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą losową prostą z rozkładu zero–jedynkowego $B_1(p)$.

Weryfikujemy układ hipotez na poziomie ufności α:

$$H_0 : p = p_0$$

$$H_1 : p < p_0$$

Statystyką testową jest:

$$Z = \frac{S_n - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$$

Test statystyczny – podsumowanie

Zbiór krytyczny dla tego układu hipotez ma postać:

$$Z < z_\alpha < 0$$

gdzie z_α jest takie, że:

$$\Phi(-z_\alpha) = 1 - \alpha \quad (\text{odczytane z tablic})$$

Wynik testu:

- jeśli $z < z_\alpha$, to **odrzucamy** H_0 i **przyjmujemy** H_1
- jeśli $z \geq z_\alpha$, to **brak podstaw do odrzucenia** H_0

Przykład banku (c.d.)

W grupie testowej 130 klientów dokładnie 5 nie spłaciło kredytu:

$$s_{130} = 5$$

Obliczamy wartość statystyki testowej:

$$z = \frac{s_{130} - 11,7}{3,26} = \frac{5 - 11,7}{3,26} \approx -2,06$$

Wartość krytyczna wynosi $z_\alpha = z_{0,05} = -1,65$

Stwierdzamy, że $z < z_\alpha$, więc należy odrzucić H_0 i przyjąć H_1.

Są powody, żeby twierdzić, że nowa procedura klasyfikowania kredytów jest lepsza od dotychczasowej.

Inne układy hipotez

Dla układu hipotez:

$$H_0 : p = p_0$$

$$H_1 : p > p_0$$

zbiór krytyczny ma postać:

$$Z > z_\alpha > 0$$

gdzie z_α jest takie, że:

$$\Phi(z_\alpha) = 1 - \alpha$$

Inne układy hipotez

Dla układu hipotez:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

zbiór krytyczny ma postać:

$$Z \in (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)$$

gdzie $z_{\alpha/2}$ jest takie, że:

$$\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

Dziękuję za uwagę!