STATYSTYKA
dr inż Krzysztof Bryś
wykład 5

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:

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

Estymacja punktowa

estymator parametru Θ - statystyka (funkcja próby), której wartość zależy od rzeczywistej wielkości parametru Θ rozkładu populacji.

estymacja punktowa - szacowanie nieznanej wartości parametru Θ na podstawie próby; polega na wyznaczeniu z próby wartości u_n estymatora U_n parametru Θ i przyjmowaniu tej wartości za oszacowanie Θ.

Estymatory wartości oczekiwanej: średnia z próby $\overline{x}$, mediana z próby $x_{0.5,n}$.

Estymatory wariancji: wariancja z próby s^2, $s_1^2 = \frac{n}{n-1}s^2$ (lepszy dla rozkładu $N(m, \sigma)$).

Estymacja przedziałowa

Przedziałem ufności dla parametru θ **na poziomie ufności** $1 - \alpha$ nazywamy przedział (θ_1, θ_2) spełniający warunki
a) θ_1, θ_2 są funkcjami próby,
b) $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$

Uwagi:
1) Przedział ufności zmienia się wraz z próbą.

2) Nieznana wartość parametru może być albo nie być w utworzonym przedziale ufności.
3) Mozna stworzyć nieskończenie wiele przedziałów ufności na danym poziomie ufności.
4) Częstość występowania prób, dla których zbudowany przedział ufności na poziomie ufności $1-\alpha$ zawiera nieznaną wartość parametru θ wynosi w przybliżeniu $1-\alpha$ (dla "dużej" liczby próbek).

Konstrukcja przedziału ufności:
1) Wybieramy estymator $U_n = U_n(\theta)$, którego rozkład dokładny lub asymptotyczny jest znany.
2) Dla danego $\alpha \in (0,1)$ dobieramy liczby a, b tak aby $P(a \leq U_n \leq b) = 1-\alpha$. (najczęściej dobieramy symetrycznie tzn. tak by $P(U_n < a) = P(U_n > b) = \frac{\alpha}{2}$)
3) Jeśli nierówność $a \leq U_n \leq b$ da się zastąpić przez $\theta_1 \leq \theta \leq \theta_2$, to przedział ufności jest postaci: (θ_1, θ_2)

Zagadnienie minimalnej liczności próby

Niech Δ-maksymalny dopuszczalny błąd oszacowania (maksymalny dopuszczalny promień przedziału ufności).

- przy szacowaniu wartości oczekiwanej m
Korzystamy z Modelu 3 (zakładaly, ze $n \geq 100$): Promień przedziału ufności$=u_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \Delta$ a zatem $n \geq (u_{1-\frac{\alpha}{2}}\sigma/\Delta)^2$

- przy szacowaniu wskaźnika struktury p **(prawdopodobieństwa sukcesu w schemacie Bernoulliego)**
Promień przedziału ufności$= u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{Z_n}{n}(1-\frac{Z_n}{n})}{n}} \leq \Delta$ a zatem $n \geq \frac{(u_{1-\frac{\alpha}{2}})^2 \cdot \frac{Z_n}{n}(1-\frac{Z_n}{n})}{\Delta^2}$.
Przypuszczalna wartość p:
$p_0 = \frac{Z_n}{n}$ jest wyznaczana z badania wstępnego (pilotażowego), szacowana na podstawie wyników poprzednich badań lub przyjmuje się $p_0 = \frac{1}{2}$.

Weryfikacja hipotez statystycznych za pomocą testów istotności.

hipoteza statystyczna- przypuszczenie dotyczące nieznanego rozkładu badanej cechy populacji.
hipoteza parametryczna- hipoteza statystyczna dotycząca wartości parametru rozkładu badanej cechy.
weryfikacja- odpowiedź na pytanie czy hipoteza statystyczna jest prawdziwa.
test statystyczny- reguła postępowania, która danej próbie przyporządkowuje decyzję przyjęcia lub odrzucenia badanej hipotezy
H_0- hipoteza zerowa (podlega badaniu)
H_1- hipoteza alternatywna
test istotności- test statystyczny, w którym wnioskowanie odbywa się przy założeniu, że hipoteza H_0 jest prawdziwa. Pozwala jedynie odrzucić H_0 (tzn. przyjąć H_1).
W przypadku weryfikacji hipotez za pomocą testów istotności wskazane jest stawianie jako H_0 hipotez co do których zachodzi podejrzenie o ich fałszywości!

Typy błędów popełnianych przy weryfikacji hipotez:
błąd 1-go rodzaju - odrzucenie prawdziwej hipotezy H_0
błąd 2-go rodzaju - przyjęcie fałszywej hipotezy H_0

poziom istotności α - prawdopodobieństwo popełnienia błędu 1-go rodzaju
β - prawdopodobieństwo popełnienia błędu 2-go rodzaju
moc testu $= 1-\beta$ - prawdopodobieństwo odrzucenia fałszywej hipotezy H_0.

Jedyny błąd jaki można popełnić weryfikując hipotezę za pomocą testu istotności to błąd 1-go rodzaju!

Zbiór krytyczny W - zbiór wartości taki, że przy założeniu, że H_0 jest prawdziwa: $P(u_n \in W) = \alpha$, gdzie u_n-obliczona wartość statystyki testowej

W praktyce $\alpha \in < 0.01; 0.1 >$.

Uwagi:

1) Przy założeniu, że H_1 prawdziwa: $P(u_n \in W) > \alpha$

2) Jeśli na poziomie istotności α_1 odrzucamy H_0, to na poziomie $\alpha_2 < \alpha_1$ może nie być podstaw do odrzucenia H_0.

Algorytm weryfikacji hipotez za pomocą testu istotności:

1. Wybieramy model.
2. Obliczamy wartość statystyki testowej u_n.
3. Budujemy zbiór krytyczny W (w zależności od postaci H_1).
4. Jeśli $u_n \in W$, to odrzucamy H_0 na poziomie istotności α. W przeciwnym przypadku mówimy, że nie ma podstaw do odrzucenia H_0.

krytyczny poziom istotności α_k - poziom: istotności, przy którym następuje zmiana decyzji weryfikacyjnej:
jeśli $\alpha < \alpha_k$ to mówimy, że nie ma podstaw do odrzucenia H_0 na poziomie istotności *alpha*
jeśli $\alpha > \alpha_k$ to odrzucamy H_0 na poziomie istotności α.

Testy zgodności

Służą do weryfikacji zgodności pomiędzy rozkładem zbioru wartości w próbie a pewnym teoretycznym rozkładem prawdopodobieństwa o dystrybuancie F_0 (gęstości prawdopodobieństwa f_0).

Weryfikowana hipoteza ma postać:
$H_0 : F = F_0$ albo $H_0 : f = f_0$
przeciw
$H_1 : F \neq F_0$ albo $H_1 f \neq f_0$,
gdzie F - nieznana dystrybuanta (f - nieznana gęstość prawdopodobieństwa) zmiennej losowej X reprezentującej badaną cechę.

Test zgodności chi-kwadrat Pearsona

Dzielimy zbiór wartości danej próby na rozłączne przedziały $I_1, \ldots, I_k$. Przy założeniu, że hipoteza H_0 jest prawdziwa,
$p_j = P(X \in I_j) = F_0(\alpha_j) - F_0(\alpha_{j-1})$, gdzie $I_j = (\alpha_{j-1}; \alpha_j)$ dla $j = 1, \ldots, k$.

Obliczamy wartość statystyki testowej:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_j - np_j)^2}{np_j},$$

gdzie n_j jest liczbą obserwacji należących do przedziału I_j, które zaobserwano w próbie, $n = \sum_{j=1}^{k} n_j$ jest liczbą wszystkich obserwacji w próbie, np_j nazywamy hipotetyczną liczbą obserwacji z przedziału I_j (jest to liczba obserwacji, które powinny należeć do I_j gdyby H_0 była prawdziwa).

Jeśli obliczona wartość statystyki χ^2 należy do zbioru krytycznego $W = (\chi^2(\alpha, k-1); +\infty)$, to odrzucamy $H_0 : F = F_0$ i przyjmujemy $H_1 : F \neq F_0$. W przeciwnym przypadku mówimy, że nie ma podstaw do odrzucenia H_0.