1 Paskaita. Atsitiktinio dydžio empirinės charakteristikos.

1.1 Empirinė pasiskirstymo funkcija, vidurkis ir dispersija

Tarkime, kad turime atsitiktinę imtį $X_1, \ldots, X_n$. Tada imties funkcija $T(X) = T(X_1, \ldots, X_n)$ yra vadinama *statistika*. Statistika $T(X)$ yra daugiamatis, arba atskiru atveju vienamatis, atsitiktinis dydis. Kai $x = (x_1, \ldots, x_n)$ yra konkreti imtis, $T(x) = T(x_1, \ldots, x_n)$ yra konkreti statistikos reikšmė, jos *realizacija*.

Viena iš pagrindinių statistikų yra vadinamoji *empirinė pasiskirstymo funkcija* $F_n(x)$. Ji apibrėžiama šitaip:

$$F_n(x) = \frac{1}{n} \sum_{X_i < x} 1,$$

kitaip tariant, tai yra mažesnių už x imties elementų X_i skaičius, padalintas iš n. Tarkime, kad $Y_1 < Y_2 < ... < Y_r$ yra skirtingi imties elementai, pasikartojantys atitinkamai $N_1, N_2, ..., N_r$ kartų, tada empirinę pasiskirstymo funkciją galime užrašyti pavidalu

$$F_n(x) = \sum_{Y_i < x} \frac{N_i}{n},$$

Apibrėšime *empirinius momentus*. Pradiniu *empiriniu l-uoju momentu* vadinsime

$$A_l = \frac{1}{n} \sum_{i=1}^{n} X_i^l.$$

Pirmasis empirinis momentas, arba *empirinis vidurkis*, paprastai žymimas $\bar{X}$

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

l-uoju empiriniu centriniu momentu vadinsime

$$M_l = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^l.$$

Antrąjį empirinį centrinį momentą, arba *empirinę dispersiją*, žymėsime

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Dydis

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

vadinamas *empiriniu standartiniu nuokrypiu*. Visi tie dydžiai yra atsitiktiniai. Konkrečioms imtims $(x_1, ..., x_n)$ gauname tų dydžių realizacijas.

1 lentelė: Pavyzdžio duomenys

7,38	7,29	7,43	7,40	7,36	7,41	7,35	7,31	7,26	7,37
7,28	7,37	7,36	7,35	7,24	7,33	7,42	7,36	7,39	7,35
7,45	7,36	7,42	7,40	7,28	7,38	7,25	7,34	7,33	7,32
7,33	7,30	7,32	7,30	7,39	7,34	7,38	7,39	7,27	7,35
7,35	7,32	7,35	7,27	7,34	7,32	7,38	7,41	7,36	7,44
7,32	7,37	7,31	7,46	7,35	7,35	7,29	7,34	7,30	7,40

2 lentelė: Sugrupuoti pavyzdžio duomenys

τ_l	7,26	7,30	7,34	7,38	7,42	7,46
n_l	5	9	20	14	9	3

1.2 Stebėjimo duomenų grupavimas

Kai stebėjimo duomenų daug, juos apdoroti gana sunku. Skaičiavimams palengvinti stebėjimo duomenys paprastai apvalinami ir grupuojami. Intervalas, kuriame telpa stebėjimo duomenys $(x_1, ..., x_n)$, paprastai skaidomas į $[\tau_l - h/2, \tau_l + h/2]$, pakeičiami skaičiais τ_l. Kuo skaičius h bus didesnis, tuo paprastesni skaičiavimai, bet tuo didesnė bus padaryta paklaida. Ir atvirkščiai, kuo mažesnis h, tuo skaičiavimai sudėtingesni, bet paklaida mažesnė.

Panagrinėsime, kaip apskaičiuojami empiriniai momentai, naudojantis sugrupuotais duomenimis. Pradinius momentus, gautus iš sugrupuotų duomenų, žymėsime a'_j, centrinius momentus – m'_j. Tarkime, kad intervale $[\tau_l - h/2, \tau_l + h/2]$ yra n_l stebėjimo duomenų. Tada

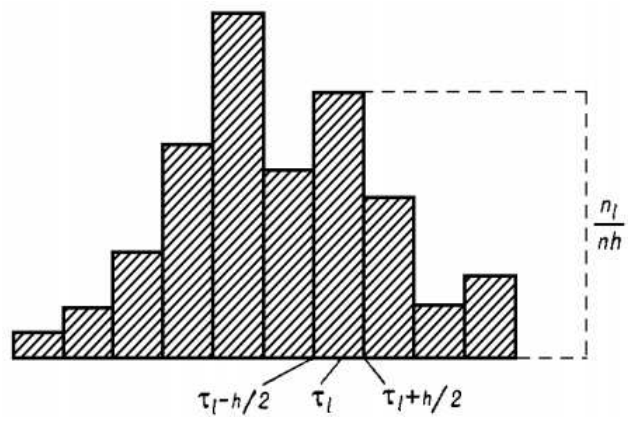
$$a'_j = \frac{1}{n} \sum_l n_l \tau_l^j,$$

sumuojame pagal visus intervalus, kuriuose yra stebėjimo duomenų. Analogiškai

$$m'_j = \frac{1}{n} \sum_l n_l (\tau_l - a'_1)^j,$$

Pavyzdys. Automatinės staklės gamina rutuliukus. Iš vienos dienos produkcijos parinkome 60 rutuliukų ir išmatavome jų skersmenis. Matavimų duomenys (milimetrais) pateikti 1 lentelėje. Po gana varginančių skaičiavimų galime gauti $a_1 = \bar{x} = 7,3490; a_2 \approx 54,10267; m_2 = s^2 \approx 0,002466$.
Skaičiavimus palengvinsime, sugrupavę duomenis. Grupavimo intervalo ilgį imsime lygų 0,04. Pasirinktus skaičius τ_l ir n_l surašysime 2 lentelėje. Vėl apskaičiuosime pirmuosius momentus. Skaičiavimai bus trumpesni. Gausime $a'_1 = 7,354667; a'_2 \approx 54,09373; m'_2 \approx 0,002612$. Kaip matome, gautos reikšmės nedaug skiriasi nuo anksčiau apskaičiuotųjų.

1 pav. pavaizduota grupuotų duomenų histograma.



1 pav.: Grupuotų duomenų histograma.

1.3 Pakankamosios statistikos

Iš visų galimų statistikų išskirsime labai svarbią jų klasę – pakankamąsias statistikas. Šią sąvoką įvedė R. Fišeris. Pakankamosios statistikos sąvoką paaiškinsime paprastu pavyzdžiu. Sakykime, turime n nepriklausomų eksperimentų. Po kiekvieno eksperimento gali įvykti kuris nors įvykis su nežinoma tikimybe p (Bernulio eksperimentai). Tarkime, kad $X_k = 1$, kai tas įvykis įvyko k-ojo eksperimento metu, ir $X_k = 0$, kai jis neįvyko. Imtis $(X_1, ..., X_n)$ rodo skaičių $T = X_1 + ... + X_n$ ir numerius eksperimentų, kai stebimasis įvykis įvyko. Intuityviai aišku, kad tų numerių žinojimas neduoda jokios papildomos informacijos apie k preikšmę. Tai galima paaiškinti ir šitaip. Imkime tokius skaičius $x_k (k = 1, \ldots, n)$, kad būtų $x_1 + ... + x_n = t$. Sąlyginis $(X_1, ..., X_n)$ skirstinys, kai $T = t$

$$P(X_1 = k_1, \ldots, X_n = k_n | T = t) = \frac{1}{C_n^t}$$

nepriklauso nuo p. Galime laikyti statistiką $T = X_1 + ... + X_n$ pakankama parametrui p įvertinti.

Apibrėžimas. *Tegu $\mathbf{X}^n = (X_1, ..., X_n)$, yra paprastoji atsitiktinė imtis. Atsitiktinio dydžio X pasiskirstymo funkcija $F_X(x) \in \mathcal{F}(\Theta)$, o Θ yra k-matė sritis ($\Theta \in \mathbf{R}^k$). Bet kuri mačioji imties $\mathbf{X}^n$ funkcija vadinama statistika.*

Apibrėžimas. *Statistika $T = T(\mathbf{X}^n), T : \mathbf{R}^n \to \mathbf{R}^k$, vadinama pakankama (nežinomam parametrui $\theta \in \Theta$), jeigu sąlyginis $\mathbf{X}^n$ skirstinys, kai žinoma statistikos T reikšmė, nepriklauso nuo nežinomo parametro θ.*

T.y., pakankamoji statistika $T = T(\mathbf{X}^n)$ turi tą pačią informaciją apie nežinomą parametrą θ, kaip ir visi stebėjimo duomenys $\mathbf{X}^n$. Yra ir kitas pakankamosios statistikos apibrėžimas.

Apibrėžimas. *Statistika $T = T(\mathbf{X}^n), T : \mathbf{R}^n \to \mathbf{R}^k$, vadinama pakankama (nežinomam parametrui $\theta \in \Theta$), jeigu egzistuoja tokia mačioji funkcija g_θ ir*

tokia Borelio funkcija h, nepriklausanti nuo θ

$$p_\theta(x) = g_\theta(T(x))h(x).$$

Pavyzdys. Stebimasis atsitiktinis dydis įgyja reikšmę 1 su nežinoma tikimybe $\alpha, 0 < \alpha < 1$ ir reikšmę 0 su tikimybe $1-\alpha$. Tada visiems $x_k, k = 1, \ldots, n$, lygiems 0 arba 1

$$p_\alpha(x_1, \ldots, x_n) = \alpha^{x_1+\ldots+x_n}(1-\alpha)^{n-(x_1+\ldots+x_n)} = (1-\alpha)^n \left(\frac{\alpha}{1-\alpha}\right)^{x_1+\ldots+x_n}.$$

Statistika $X_1 + ... + X_n$ yra pakankama parametrui α.

Pavyzdys. Stebimasis Puasono atsitiktinis dydis su nežinomu parametru $\lambda > 0$, tada visiems sveikiesiems neneigiamiems $x_k, k = 1, \ldots, n$,

$$p_\lambda(x_1, \ldots, x_n) = \frac{\lambda^{x_1+\ldots+x_n}}{x_1! \cdot \ldots \cdot x_n!} e^{-\lambda n}.$$

Statistika $X_1 + ... + X_n$ ir čia yra pakankama parametrui λ.

Pavyzdys. Imkime normalųjį dėsnį $N(a, \sigma^2)$. Atsitiktinė imtis $(X_1, \ldots, X_n)$ turi tankį

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - a)^2\}. \tag{1}$$

Jei a žinomas, o $\sigma > 0$ – nežinomas, tai statistika

$$\sum_{k=1}^{n}(X_k - a)^2$$

yra pakankama parametrui σ^2. Kai σ žinomas, o $a \in \mathbf{R}$ – nežinomas, užrašę (1)

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\{-\frac{na^2}{2\sigma^2} + \frac{a}{\sigma^2}\sum_{k=1}^{n}x_k\} \cdot exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}x_k^2\}, \tag{2}$$

matome, kad statistika $X_1 + ... + X_n$ yra pakankama parametrui a. Jei abu parametrai $a \in \mathbf{R}$ ir $\sigma > 0$ yra nežinomi, tai iš (2) išplaukia, kad vektorinė statistika (T_1, T_2), kur

$$T_1 = X_1 + ... + X_n, \quad T_2 = X_1^2 + ... + X_n^2$$

yra pakankama parametrams (a, σ^2).