11 Paskaita. Atsitiktinių dydžių kovariacija ir koreliacijos koeficientas

11.1 Apibrėžimai ir savybės

Apibrėžimas. *Tegu atsitiktinių dydžių ξ_1, ξ_2 antrieji momentai egzistuoja. Dydį*

$$cov(\xi_1, \xi_2) = \mathbf{E}(\xi_1 - \mathbf{E}\xi_1)(\xi_2 - \mathbf{E}\xi_2)$$

vadinsime dydžių ξ_1, ξ_2 kovariacija, o dydį

$$\rho(\xi_1, \xi_2) = \frac{cov(\xi_1, \xi_2)}{\sqrt{\mathbf{D}\xi_1 \mathbf{D}\xi_2}},$$

kai vardiklis nelygus nuliui – koreliacijos koeficientu. Jei bent vienas iš dydžių ξ_1, ξ_2 yra išsigimęs, tai sakysime, kad koreliacijos koeficientas lygus nuliui.

Nesunku įsitikinti, kad kovariaciją galima skaičiuoti ir taip:

$$cov(\xi_1, \xi_2) = \mathbf{E}\xi_1\xi_2 - \mathbf{E}\xi_1 \mathbf{E}\xi_2.$$

Pavyzdys. Skaičiais pažymėti rutuliai

Urnoje yra trys balti rutuliai, pažymėti skaičiais 1, 0, 0 ir du juodi, ant kurių užrašyti skaičiai 1, 1. Atsitiktinai be grąžinimo traukiami du rutuliai, dydis X lygus baltų rutulių skaičiui, o Y – skaičių, užrašytų ant rutulių, sumai. Apskaičiuosime dydžių kovariaciją. Iš pradžių sudarykime tikimybių $P(X = x, Y = y)$ lentelę

	$X=0$	$X=1$	$X=2$	
$Y=0$	0	0	0,1	0,1
$Y=1$	0	0,4	0,2	0,6
$Y=2$	0,1	0,2	0	0,3
	0,1	0,6	0,3	

Pavyzdžiui, $P(X = 2, Y = 0) = \dfrac{C_2^2}{C_5^2}$. Susumavę skaičius surašytus stulpeliuose, gauname dydžio X reikšmių tikimybes, o eilutėse – dydžio Y reikšmių tikimybes. Skaičiuodami vidurkius galime nerašyti tų dėmenų, kurie lygūs nuliui:

$$\begin{aligned}
\mathbf{E}X &= 1 \cdot 0,6 + 2 \cdot 0,3 = 1,2,\\
\mathbf{E}Y &= 1 \cdot 0,6 + 2 \cdot 0,3 = 1,2,\\
\mathbf{E}XY &= 1 \cdot 1 \cdot 0,4 + 1 \cdot 2 \cdot 0,2 + 2 \cdot 1 \cdot 0,2 = 1,2,\\
cov(X,Y) &= 1,2 - 1,2 \cdot 1,2 = -0,24.
\end{aligned}$$

Pasvarstykime, kodėl pavyzdžio dydžiams gavome neigiamą kovariacijos reikšmę. Kai dydis X įgyja didesnes už vidurkį reikšmes, t.y. $X - \mathbf{E}X > 0$, tai dydis Y linkęs įgyti mažesnes reikšmes, t.y. $Y - \mathbf{E}Y < 0$ ir atvirkščiai. Taip ir yra: daugiau baltų rutulių – mažesnė skaičių suma, nes baltųjų rutulių numeriai mažesni. Taigi jei kovariacija yra neigiama, tai esant didesnėms vieno dydžio reikšmėms, kito dydžio reikšmės linkusios būti mažesnės. Jeigu kovariacija yra teigiama, tai esant didesnėms vieno dydžio reikšmėms ir kito dydžio reikšmės linkusios būti didesnės.
Apskaičiuosime koreliacijos koeficientą $\rho(X, Y)$.

$$\begin{aligned}\mathbf{E}X^2 = \mathbf{E}Y^2 &= 1 \cdot 0,6 + 4 \cdot 0,3 = 1,8,\\ \mathbf{D}X = \mathbf{D}Y &= 1,8 - 1,2^2 = 0,36,\\ \rho(X,Y) &= \frac{-0,24}{\sqrt{0,36 \cdot 0.36}} = -0,667.\end{aligned}$$

Apibrėžimas. *Jeigu X, Y yra atsitiktiniai dydžiai ir $cov(X, Y) > 0$, tai dydžius vadinsime teigiamai koreliuotais, jeigu $cov(X, Y) < 0$, dydžius vadinsime neigiamai koreliuotais. Jeigu $cov(X, Y) = 0$ dydžius vadinsime nekoreliuotais.*

Teorema. *Tegu ξ_1, ξ_2 yra atsitiktiniai dydžiai, turintys baigtines ir nenulines dispersijas. Teisingi tokie teiginiai:*

1. $|\rho(\xi_1, \xi_2)| \leq 1$;
2. $\rho(\xi_1, \xi_2) = 0$ *tada ir tik tada, kai* $\mathbf{D}[\xi_1 + \xi_2] = \mathbf{D}\xi_1 + \mathbf{D}\xi_2$;
3. *lygybė* $|\rho(\xi_1, \xi_2)| = 1$ *teisinga tada ir tik tada, kai egzistuoja konstantos $\lambda_1, \lambda_2, \lambda_3$, ne visos lygios nuliui, kad*

$$P(\lambda_1\xi_1 + \lambda_2\xi_2 = \lambda_3) = 1.$$

Įrodymas.

1. Įrodymas išplaukia iš koreliacijos koeficiento apibrėžimo ir šios nelygybės momentams:
$$\mathbf{E}|\xi_1\xi_2| \leq (\mathbf{E}\xi_1)^{1/2}(\mathbf{E}\xi_2)^{1/2},$$
kurioje reikia vietoje ξ_i įrašyti $\mathbf{E}[\xi_i - \mathbf{E}\xi_i], i = 1, 2$.

2. Imdami vidurkius abiejose tapatybės
$$(\xi_1 + \xi_2 - \mathbf{E}\xi_1 - \mathbf{E}\xi_2)^2 = (\xi_1 - \mathbf{E}\xi_1)^2 + (\xi_2 - \mathbf{E}\xi_2)^2 + 2(\xi_1 - \mathbf{E}\xi_1)(\xi_2 - \mathbf{E}\xi_2)$$
pusėse gausime
$$\mathbf{D}[\xi_1 + \xi_2] = \mathbf{D}\xi_1 + 2cov(\xi_1, \xi_2) + \mathbf{D}\xi_2.$$
Iš čia išplaukia 2-as teiginys.

3. Tegu
$$P(\lambda_1\xi_1 + \lambda_2\xi_2 = \lambda_3) = 1$$
ir $\lambda_1 \neq 0$. Tada $P(\xi_1 = a\xi_2 + b) = 1$. Iš čia
$$\mathbf{D}\xi_1 = a^2\mathbf{D}\xi_2, \quad \xi_1 - \mathbf{E}\xi_1 = a(\xi_2 - \mathbf{E}\xi_2), \quad cov(\xi_1, \xi_2) = a\mathbf{D}\xi_2.$$
Tada
$$\rho(\xi_1, \xi_2) = \frac{cov(\xi_1, \xi_2)}{\sqrt{\mathbf{D}\xi_1\mathbf{D}\xi_2}} = \frac{a}{|a|} = \operatorname{sgn} a,$$
Tegu dabar $\rho(\xi_1, \xi_2) = 1$. Apibrėžkime atsitiktinį dydį
$$\eta = \left(\frac{\xi_1 - \mathbf{E}\xi_1}{\sqrt{\mathbf{D}\xi_1}} - \frac{\xi_2 - \mathbf{E}\xi_2}{\sqrt{\mathbf{D}\xi_2}}\right)^2 \geq 0.$$
Nesunku patikrinti, kad $\mathbf{E}\eta = 0$. Bet tada turi būti
$$P\left(\frac{\xi_1 - \mathbf{E}\xi_1}{\sqrt{\mathbf{D}\xi_1}} - \frac{\xi_2 - \mathbf{E}\xi_2}{\sqrt{\mathbf{D}\xi_2}} = 0\right) = 1.$$
Analogiškai tiriame atvejį $\rho(\xi_1, \xi_2) = -1$. Taigi, radome konstantas $\lambda_1, \lambda_2, \lambda_3$, ne visos lygios nuliui, kad
$$P(\lambda_1\xi_1 + \lambda_2\xi_2 = \lambda_3) = 1.$$

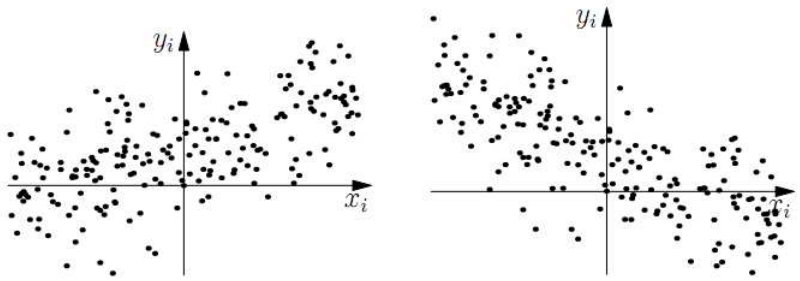
Koreliacijos koeficientas nepriklauso nuo to, kokie matavimo vienetai naudojami dydžių reikšmėms užrašyti. Tai išplaukia iš teoremos.

Teorema. *Tegu X, Y yra atsitiktiniai dydžiai, turintys teigiamas dispersijas, o a_1, a_2, b_1, b_2 – bet kokie skaičiai, $a_1 \cdot a_2 \neq 0$. Tada*
$$\rho(a_1X + b_1, a_2Y + b_2) = \begin{cases} \rho(X, Y), & \text{jei } a_1 \cdot a_2 > 0, \\ -\rho(X, Y), & \text{jei } a_1 \cdot a_2 < 0. \end{cases}$$

Taigi koreliacijos koeficientas nesikeičia, kai dydžius tiesiškai transformuojame su to paties ženklo "skalės keitimo" koeficientais a_1, a_2.

Tarkime pakartoję bandymą n kartų, gavome atsitiktinių dydžių X, Y reikšmių poras
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$
Jas galime pavaizduoti plokštumos taškais. Jeigu dydžiai būtų teigiamai koreliuoti, tai taškai išsidėstytų pagal tiesę su teigiamu krypties koeficientu, žr. 1 pav. Jeigu neigiamai koreliuoti – taškai susitelktų apie tiesę su neigiamu krypties koeficientu.

Jeigu $\rho(X, Y) = 0$, taškai sudarytų "debesį" ir grupavimosi apie jokią tiesę negalėtume įžvelgti. Taigi koreliacijos koeficientas naudojamas dviejų dydžių tiesinės priklausomybės laipsniui matuoti.



1 pav.: Teigiamai ir neigiamai koreliuotų atsitiktinių dydžių reikšmių vaizdavimas plokštumoje.

11.2 Koreliacijos koeficiento savybės

Teorema. *Tegu X, Y yra atsitiktiniai dydžiai, turintys baigtines ir nenulines dispersijas. Teisingi tokie teiginiai:*

1. $-1 \leqslant \rho(X,Y) \leqslant 1$;
2. *Jei $Y = aX + b$ kur a ir b yra skaičiai, tai $\rho(X,Y) = 1$, jei $a > 0$ ir $\rho(X,Y) = -1$, jei $a < 0$;*
3. *Jei $\rho(X,Y) = \pm 1$ tai egzistuoja skaičiai $a \neq 0, b$*

$$P(Y = aX + b) = 1.$$

11.3 Nepriklausomų atsitiktinių dydžių sumos skirstinys

Dažnai tenka nagrinėti dviejų ar daugiau nepriklausomų atsitiktinių dydžių sumą. Pavyzdžiui, dalyvavus dvejose loterijose natūralu suvesti abiejų laimėjimų arba pralaimėjimų balansą. Pavyzdžiui, binominį atsitiktinį dydį $X \sim \mathcal{B}(n,p)$ reiškiame Bernulio atsitiktinių dydžių X_i suma:

$$X = X_1 + X_2 + \ldots + X_n,$$

čia $X_i = 1$, jei i-me bandyme pasirodė sėkmė, $X_i = 0$, jei buvo nesėkmė, o p yra sėkmės tikimybė kiekviename bandyme. Šie dydžiai susiję su nepriklausomais bandymais, taigi ir patys yra nepriklausomi. Dviejų atsitiktinių dydžių suma taip pat yra atsitiktinis dydis. Kaip jo skirstinys priklauso nuo dėmėnų skirstinių?

Teorema. *Jei ξ, η yra nepriklausomi atsitiktiniai dydžiai, $\zeta = \xi + \eta$, tai*

$$P(\zeta = z) = \sum_{x,y:x+y=z} P(\xi = x)P(\eta = y). \tag{1}$$

Pavyzdys. Tegu X ir Y yra nepriklausomi binominiai atsitiktiniai dydžiai su parametrais (n,p) ir (m,p). Rasime atsitiktinio dydžio $X+Y$ skirstinį. Pasinaudoję formule (1) bei X ir Y nepriklausomumu gausime:

$$\begin{aligned} P(X+Y=k) &= \sum_{i=0}^{n} P(X=i, Y=k-i) = \sum_{i=0}^{n} P(X=i)P(Y=k-i) \\ &= \sum_{i=0}^{n} C_n^i p^i q^{n-i} C_m^{k-i} p^{k-i} q^{m-k+i} = p^k q^{n+m-k} \sum_{i=0}^{n} C_n^i C_m^{k-i} \\ &= C_{n+m}^k p^k q^{n+m-k}. \end{aligned}$$

Matome, kad atsitiktinio dydžio $X+Y$ skirstinys yra binominis su parametrais $(n+m,p)$. Tai visiškai natūralu, nes $X+Y$ yra atsitiktinis dydis, kuris reiškia sėkmių skaičių atlikus $n+m$ nepriklausomų eksperimentų, kiekviename iš kurių sėkmės tikimybė yra lygi p.
Atsakysime į klausimą apie nepriklausomų atsitiktinių dydžių sumos skirstinį, kai dėmėnys yra tolydieji atsitiktiniai dydžiai.

Teorema. *Jei ξ_1, ξ_2 yra absoliučiai tolydieji nepriklausomi atsitiktiniai dydžiai, turintys tankius p_{ξ_1}, p_{ξ_2}, tai atsitiktinis dydis $\eta = \xi_1 + \xi_2$ yra taip pat absoliučiai tolydusis ir jo tankis lygus*

$$p_{\xi_1+\xi_2}(u) = \int_{-\infty}^{\infty} p_{\xi_1}(v) p_{\xi_2}(u-v)dv = \int_{-\infty}^{\infty} p_{\xi_2}(v) p_{\xi_1}(u-v)dv. \qquad (2)$$

Pavyzdys. Tegu X ir Y yra nepriklausomi atsitiktiniai dydžiai, turintys tolygųjį skirstinį intervale $(0,1)$. Rasime atsitiktinio dydžio $X+Y$ skirstinį. Iš lygties (2) ir tankio funkcijų

$$p_X(u) = p_Y(u) = \begin{cases} 1, & 0 < u < 1, \\ 0, & \text{kitur} \end{cases}$$

gauname

$$p_{X+Y}(u) = \int_0^1 p_X(u-v) \cdot 1 dv.$$

Kai $0 \leqslant u \leqslant 1$,

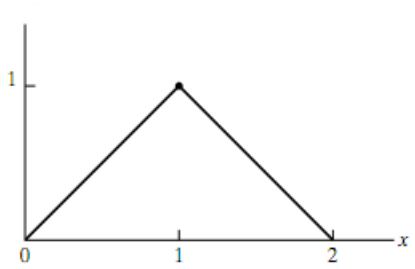
$$p_{X+Y}(u) = \int_0^u dv = u.$$

Kai $1 < u < 2$,

$$p_{X+Y}(u) = \int_{u-1}^1 dv = 2-u.$$

Taigi

$$p_{X+Y}(u) = \begin{cases} u, & 0 \leqslant u \leqslant 1, \\ 2-u, & 1 < u < 2, \\ 0, & \text{kitur.} \end{cases}$$



2 pav.: Trikampio skirstinio tankio funkcija.

Atsitiktinis dydis $X + Y$ turi trikampį skirstinį, nes jo tankio funkcija turi trikampio formą (žr. 2 pav.)

Teorema. *Jei $X_i, i = 1, 2, \ldots, n$ yra nepriklausomi normaliai pasiskirstę atsitiktiniai dydžiai, turintys vidurkius μ_i ir dispersijas $\sigma_i^2, i = 1, 2, \ldots, n$, tai jų suma $\sum_{i=1}^n X_i$ yra taip pat normaliai pasiskirstęs atsitiktinis dydis su parametrais $\mu = \sum_{i=1}^n \mu_i$ ir $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.*