5. Závislé pravděpodobnosti

In: Bohuslav Hostinský (author): Počet pravděpodobnosti. Druhá část. (Czech). Praha: Jednota československých matematiků a fysiků, 1950. pp. 3–21.

Persistent URL: http://dml.cz/dmlcz/403304

ZÁVISLÉ PRAVDĚPODOBNOSTI

48. Podmíněné pravděpodobnosti. Zjev A může se objeviti jakožto výsledek nějakého pokusu, který se koná za daných podmínek; vzhledem k nim má zjev A určitou *prostou pravděpodobnost* $P(A)$. *Podmíněná pravděpodobnost* zjevu A za předpokladu, že nastal jiný zjev B, značí se $P_B(A)$ a liší se obecně od $P(A)$. Obdobně zavádíme prostou pravděpodobnost $P(B)$ zjevu B a jeho podmíněnou pravděpodobnost $P_A(B)$ za předpokladu, že nastal A.

Nastane-li zjev A, nemá to vlivu na hodnotu pravděpodobnosti $P_A(B)$; nastane-li B za předpokladu, že nastal**) A, nemá to vlivu na hodnotu pravděpodobnosti $P(A)$. Proto prostá pravděpodobnost $P(A, B)$, že nastanou oba zjevy A a B, rovná se podle pravidla o násobení pravděpodobností součinu $P(A)\,P_A(B)$. V tomto součinu lze zaměniti A s B, takže

$$P(A, B) = P(A)\,P_A(B) = P(B)\,P_B(A). \qquad (1)$$

Rovnice (1) udává souvislost mezi prostou pravděpodobností $P(A, B)$, že nastanou zjevy A a B, prostými pravděpodobnostmi $P(A)$, $P(B)$ obou zjevů a podmíněnými pravděpodobnostmi $P_A(B)$ a $P_B(A)$.

Obecně je

$$P(A, B) \neq P(A)\,P(B).$$

Ve zvláštním případě, že A a B jsou zjevy vzájemně nezávislé, je

$$P_B(A) = P(A),\ P_A(B) = P(B);$$

rovnice (1) se pak redukují na jedinou:

*) Číslování kapitol a odstavců navazuje na první část, která vyšla v Cestě k vědění, sv. 53.

**) Není nutno, aby zjev B časově následoval po zjevu A; B může býti současný s A nebo může nastati i před zjevem A.

$$P(A, B) = P(A)\, P(B),$$

která vyjadřuje pravidlo o složené pravděpodobnosti (odst. 5).

49. Příklady podmíněných pravděpodobností. a) Jsou dána dvě osudí, jedno bílé, které obsahuje dvě bílé koule a jednu černou, druhé pak černé, které obsahuje dvě černé koule a jednu bílou. Volíme jedno osudí (pravděpodobnost voliti bílé je $\frac{1}{2}$, pro černé též $\frac{1}{2}$) a vytáhneme kouli, kterou vložíme zpět; byla-li to bílá, konáme druhý tah z bílého osudí, byla-li to černá, konáme jej z černého. Zjev A nechť je vytažení bílé koule při prvním tahu, zjev B vytažení bílé koule při druhém tahu. Jak veliké jsou prosté pravděpodobnosti $P(A)$, $P(B)$ a jak veliká je pravděpodobnost $P(A, B)$, že při prvním i při druhém tahu vyjde bílá?

$P(A)$ se rovná součtu dvou složených pravděpodobností; buď volíme bílé osudí a vytáhneme bílou (pravděpodobnost $\frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$) nebo volíme černé a vytáhneme bílou (pravděpodobnost $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$) a tedy

$$P(A) = \tfrac{1}{3} + \tfrac{1}{6} = \tfrac{1}{2}.$$

Pravděpodobnost vytáhnouti při prvním tahu černou jest ovšem také $\frac{1}{2}$.

$P(B)$ je rovněž součet dvou složených pravděpodobností: buď v prvním tahu vyjde bílá a při druhém vytáhneme z bílého osudí bílou (pravděpodobnost $\frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$) nebo v prvním tahu vyjde černá a při druhém vytáhneme z černého osudí bílou (pravděpodobnost $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$). Je tedy

$$P(B) = \tfrac{1}{3} + \tfrac{1}{6} = \tfrac{1}{2}.$$

Je-li známo, že v prvním tahu vyšla bílá, koná se druhý tah z bílého osudí, a tedy $P_A(B) = \frac{2}{3}$ je pravděpodobnost, že i ve druhém vyjde bílá. Prostou pravděpodobnost $P(A, B)$ vypočteme dvojím způsobem:

1. Podle rovnice (1) odst. 48 je

$$P(A, B) = P(A)\, P_A(B) = \tfrac{1}{2} \cdot \tfrac{2}{3} = \tfrac{1}{3}.$$

2. Pravděpodobnost p_1, voliti bílé osudí, vytáhnouti bílou a pak znova z bílého vytáhnouti bílou, je $p_1 = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{2}{9}$; pravděpodobnost p_2 voliti černé osudí, vytáhnouti bílou a pak z bílého vytáhnouti bílou, je $p_2 = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{9}$ a

$$P(A, B) = p_1 + p_2 = \frac{2}{9} + \frac{1}{9} = \frac{1}{3}.$$

Je tedy

$$P(A)\,P(B) = \frac{1}{4},\ P(A, B) = \frac{1}{3},\ P(A, B) > P(A)\,P(B).$$

Pravděpodobnost $P_B(A)$, že v prvním tahu vyšla bílá, je-li známo, že v druhém tahu vyšla bílá, vypočte se podle rovnice (1), odst. 48:

$$P_B(A) = \frac{P(A, B)}{P(B)} = \frac{1}{3} : \frac{1}{2} = \frac{2}{3}.$$

b) Pro obyvatele určitého města rozeznáváme několik pravděpodobností vztahujících se ke zjevům: A přijíti do nemocnice, a B zemříti.

Pravděpodobnost, že někdo během 1 roku zemře ..... $P(B) = 0{,}006$

Pravděpodobnost, že někdo během 1 roku přijde do nemocnice ..... $P(A) = 0{,}02$

Pravděpodobnost, že někdo, přijde-li během 1 roku do nemocnice, zemře tam během téhož roku..... $P_A(B) = 0{,}06$

Zde uvedené číselné hodnoty prostých pravděpodobností $P(A)$, $P(B)$ a podmíněné pravděpodobnosti $P_A(B)$ vyjadřují tyto poměry: má-li město 100 000 obyvatelů, zemře z nich během roku celkem 600; do nemocnice přijde během roku 2000 osob, z nichž 120 tam zemře během téhož roku.

Podle rovnic odst. 48 je pravděpodobnost, že osoba, která zemřela během roku, zemřela v nemocnici, rovna

$$P_B(A) = \frac{P_A \cdot P_A(B)}{P(B)} = \frac{0{,}02 \cdot 0{,}06}{0{,}006} = 0{,}2.$$

Pravděpodobnost, že někdo přijde do nemocnice a že tam zemře, je

$$P(A, B) = P(A)\, P_A(B) = 0{,}02 \,.\, 0{,}06 = 0{,}0012.$$

Je tedy

$$P(A)\, P(B) = 0{,}00012,\ P(A, B) = 0{,}0012,$$
$$P(A, B) > P(A)\, P(B).$$

Úmrtnost v nemocnici, vyjádřená pravděpodobností $P_A(B)$, je v našem případě desetkrát větší než prostá úmrtnost v městě vyjádřená pravděpodobností $P(B)$ (podle *S. Bernštejna*).

50. Tabulky úmrtnosti. a) Číselné vyjádření úmrtnosti odvozuje se ze změn, které nastávají v souboru osob žijících za celkem stejných podmínek. Budiž dán t. zv. *základní soubor* složený z l_0 osob, které se narodily v témže roce. Z těchto l_0 osob je po uplynutí x let na živu l_x osob; $l_0 - l_x$ osob tedy zemřelo během x let. Z toho plyne: pravděpodobnost, že někdo se dožije věku x let, je rovna

$$\frac{l_x}{l_0}. \tag{1}$$

Pravděpodobnost, že někdo zemře dříve, než dosáhne věku x, je

$$1 - \frac{l_x}{l_0} = \frac{l_0 - l_x}{l_0}. \tag{2}$$

b) Pravděpodobnost p_{xy}, že x-letá osoba dožije se věku y ($x < y$), počítá se obdobně jako pravděpodobnost (1) s tím rozdílem, že základní soubor je zde utvořen l_x osobami x-letými; platí

$$p_{xy} = \frac{l_y}{l_x}. \tag{3}$$

Pravděpodobnost, že x-letá osoba se nedožije stáří y, je

$$1 - p_{xy} = \frac{l_x - l_y}{l_x}. \tag{4}$$

Pravděpodobnost p_{xy} dožíti se stáří y let je pravděpodobnost závislá, neboť závisí na stáří x, kterého osoba již dosáhla.

c) Pravděpodobnost, že a-letá osoba zemře do roka, je podle (4) pro $x = a$, $y = a + 1$ rovna

$$\frac{l_a - l_{a+1}}{l_a}. \tag{5}$$

Pravděpodobnost, že a-letá osoba bude žíti ještě x let a že během následujícího roku zemře, je

$$\frac{l_{a+x}}{l_a} \cdot \frac{l_{a+x} - l_{a+x+1}}{l_{a+x}} = \frac{l_{a+x} - l_{a+x+1}}{l_a}. \tag{6}$$

d) Veličiny $l_0, l_1, l_2, \ldots$ jsou sestaveny v tabulkách. Ve Valouchových tabulkách*) je tabulka úmrtnosti vypracovaná Státním úřadem statistickým. V tabulce pro muže je vzato za základ číslo $l_0 = 100\,000$; tabulka končí číslem $l_{104} = 1$, l_{105} a další jsou rovny 0. Tabulka pro ženy (také se základem $l_0 = 100\,000$) končí číslem $l_{106} = 1$.

51. Podmíněné střední hodnoty. Označme písmeny $A_1, A_2, \ldots, A_r$ zjevy, které se mohou vyskytnouti jakožto výsledky nějakého pokusu; pokus vede vždy k jedinému z nich. Nechť je p_k pravděpodobnost, že zjev A_k bude výsledkem pokusu ($k = 1, 2, \ldots, r$). Platí rovnice

$$p_1 + p_2 + \ldots + p_r = 1.$$

Přiřaďme zjevu A_k veličinu α_k. Proměnná veličina x, závislá na výsledku pokusu, budiž rovna α_k, vyskytne-li se zjev A_k. Pak je *prostá střední hodnota veličiny* x rovna

$$E(x) = p_1\alpha_1 + p_2\alpha_2 + \ldots + p_r\alpha_r.$$

*) *M. Valouch* a *M. A. Valouch*: Tabulky logaritmické, 10. vydání (1937), str. 102—105.

Budiž nyní, ve shodě s označením zavedeným v odst. 48, $P_B(A_k)$ pravděpodobnost, že se zjev A_k vyskytne za předpokladu, že se vyskytl mimo to jiný zjev B; předpokládáme, že objevení se zjevu B nezávisí na výsledku shora uvažovaného pokusu vedoucího k jednomu ze zjevů $A_1, A_2, \ldots, A_r$. Pak je *podmíněná střední hodnota veličiny* x za předpokladu, že nastal zjev B, rovna

$$E_B(x) = P_B(A_1)\,\alpha_1 + P_B(A_2)\,\alpha_2 + \ldots + P_B(A_r)\,\alpha_r. \quad (1)$$

52. Příklady podmíněných středních hodnot. a) Budiž L_a *střední délka věku*, kterého se dočká a-letá osoba. Užívajíce tabulky úmrtnosti (odst. 50) vypočteme L_a takto:

Podle rovnice (6), odst. 50 je

$$\frac{l_{a+x} - l_{a+x+1}}{l_a}$$

pravděpodobnost, že a-letá osoba bude žíti ještě x let a že pak během následujícího roku zemře.

V rovnici (1) odst. 51 nechť A_k značí zjev: a-letá osoba bude žíti ještě k let a během následujícího roku zemře (ve stáří mezi $(a + k)$ a $(a + k + 1)$. V téže rovnici nechť B značí zjev: osoba dožije se stáří a. Dosaďme do rovnice (1), odst. 51:

$$\alpha_k = k, \quad P_B(A_k) = \frac{l_{a+k} - l_{a+k+1}}{l_a};$$

její pravá strana bude pak rovna střední délce zbývajícího života pro a-letou osobu, tedy $L_a - a$. Je tedy

$$L_a - a = \frac{l_{a+1} - l_{a+2}}{l_a} \cdot 1 + \frac{l_{a+2} - l_{a+3}}{l_a} \cdot 2 + \frac{l_{a+3} - l_{a+4}}{l_a} \cdot 3 + \ldots$$

nebo

$$L_a = a + \frac{l_{a+1} + l_{a+2} + l_{a+3} + \ldots}{l_a}.$$

Řada stojící v čitateli má kladné členy potud, pokud v tabulce l_x není rovno nule. Jedná-li se na př. o muže, je poslední člen řady $l_{104} = 1$; l_{105} a další jsou rovny nule. V tabulkách se někdy připojuje k napsanému zlomku hodnota $\frac{1}{2}$, čímž se vyjadřuje, že osoba se může dožíti části (průměrně poloviny) posledního (pro muže 105.) roku, ač tabulka dává pro ten rok nulovou hodnotu l_{105}.

b) Konáme řadu vzájemně nezávislých pokusů; budiž p pravděpodobnost, že se pokus zdaří, stejná pro každý pokus. Přiřadíme k-tému pokusu určitou veličinu $x^{(k)}$, ($k = 1, 2, 3, \ldots$), která se rovná 1, zdaří-li se pokus a která se rovná 0, nezdaří-li se. Užijeme označení zavedeného v odst. 13 a 14: m nechť značí skutečný počet zdařených pokusů, vykonáme-li celkem n pokusů, a h nechť je úchylka čísla m od jeho střední hodnoty np. Střední hodnotu nějakého čísla x budeme značiti $E(x)$ (dříve jsme užívali znaku s. h. (x)). Podle odst. 14b je

$$x^{(1)} + x^{(2)} + \ldots + x^{(n)} = m,$$
$$h = x^{(1)} - p + x^{(2)} - p + \ldots + x^{(n)} - p = m - np,$$
$$E(x^{(k)} - p) = 0,\ E[(x^{(k)} - p)^2] = p(1 - p),$$
$$E[(x^{(k)} - p)(x^{(l)} - p)] = 0 \text{ pro } l \neq k, \qquad (1)$$
$$E(h) = 0,\ E(h^2) = np(1 - p). \qquad (2)$$

Prodlužme nyní řadu pokusů tak, že celkový jich počet bude $N (n < N)$; příslušnou úchylku nazveme H. Bude tedy

$$H = x^{(1)} - p + x^{(2)} - p + \ldots + x^{(n)} - p + x^{(n+1)} - p + \\ + \ldots + x^{(N)} - p,\ E(H) = 0.$$

Klademe si za úlohu vypočítati střední hodnotu součinu hH ve dvou různých případech: jednak za předpokladu, že h má danou známou hodnotu, jednak prostou střední hodnotu.

Podmíněná střední hodnota součinu hH za předpokladu, že h má danou známou hodnotu h je (viz odst. 10c)

$$E_h(h \,.\, H) = h \,.\, E(H) = 0.$$

Prostá střední hodnota součinu hH je ($n < N$)

$$E(hH) = E\{[x^{(1)} - p + x^{(2)} - p + \ldots + x^{(n)} - p] \cdot [x^{(1)} - p + x^{(2)} - p + \ldots + x^{(N)} - p]\} =$$
$$= E[(x^{(1)} - p)^2 + (x^{(2)} - p)^2 + \ldots + (x^{(n)} - p)^2] =$$
$$= np(1 - p) = E(h^2),$$

neboť podle (1) má každý čtverec $(x^{(k)} - p)^2$ střední hodnotu $p(1 - p)$ a každý součin $(x^{(k)} - p)(x^{(l)} - p)$ střední hodnotu rovnou nule pro $k \neq l$. Výsledek shrneme takto:

Budiž H úchylka v řadě složené z N nezávislých pokusů a h úchylka v řadě složené z n prvních pokusů ($n < N$). *Podmíněná střední hodnota součinu hH za předpokladu, že h má známou hodnotu, se rovná nule. Prostá střední hodnota součinu hH se rovná střední hodnotě čtverce úchylky h a nezávisí na čísle N.**)

53. Jak se normalisuje veličina závislá na náhodě. Budiž x veličina závislá na náhodě. V některých úlohách se hodí *normalisovati* veličinu x, t. j. zavésti do počtu novou veličinu ξ, která je lineární funkcí veličiny x a která má vlastnosti:

$$E(\xi) = 0, \quad E(\xi^2) = 1. \tag{1}$$

Veličinu ξ odvodíme z x takto: odečteme od x její střední hodnotu $E(x)$ a rozdíl dělíme odmocninou ze střední hodnoty čtverce veličiny $[x - E(x)]$. Je tedy

$$\xi = \frac{x - E(x)}{\sqrt{E[x - E(x)]^2}}. \tag{2}$$

Výraz na pravé straně rovnice (2) má vlastnosti vyjádřené rovnicemi (1). Veličina ξ je *normalisovaná veličina závislá na náhodě*. Střední hodnota nalézající se pod znamením odmocniny v (2) může se upraviti takto:

$$E[x - E(x)]^2 = E(x^2) - 2[E(x)]^2 + [E(x)]^2 =$$
$$= E(x^2) - [E(x)]^2. \tag{3}$$

*) Viz *P. Lévy*: Commentarii Math. Helvetici, vol. *16* (1943—44), pg. 242.

Místo (2) dostaneme, užijeme-li této úpravy,

$$\xi = \frac{x - E(x)}{\sqrt{E(x^2) - [E(x)]^2}}. \tag{4}$$

54. Korelace a koeficient korelace. a) K pojmu korelace docházíme sledujíce souvislost dvou znaků na nějakém jedinci nebo vůbec dvou proměnných veličin x a y, které v různých případech současně nabývají různých hodnot. Nechť jsou $x_1, x_2, \ldots, x_i, \ldots$ možné hodnoty veličiny x a $y_1, y_2, \ldots, y_k, \ldots$ možné hodnoty veličiny y. Jeden „případ" jest určen, známe-li příslušný pár hodnot x_i a y_k.

Budiž $P(x_i, y_k)$ prostá pravděpodobnost případu, ve kterém $x = x_i$ a zároveň $y = y_k$. Podmíněné pravděpodobnosti:

$P_{y_k}(x_i)$, že $x = x_i$, dáno-li, že $y = y_k$, a
$P_{x_i}(y_k)$, že $y = y_k$, dáno-li, že $x = x_i$

souvisí s prostými pravděpodobnostmi

$$P(x_i), \text{ že } x = x_i;\ P(y_k), \text{ že } y = y_k$$

podle rovnice (1), odst. 48; tato rovnice má nyní tvar

$$P(x_i, y_k) = P(x_i)\, P_{x_i}(y_k) = P(y_k)\, P_{y_k}(x_i).$$

Poněvadž pak

$$P(x_i) = \sum_k P(x_i, y_k), \qquad P(y_k) = \sum_i P(x_i, y_k), \tag{1}$$

je

$$P_{x_i}(y_k) = \frac{P(x_i, y_k)}{\sum_k P(x_i, y_k)}, \quad P_{y_k}(x_i) = \frac{P(x_i, y_k)}{\sum_i P(x_i, y_k)}. \tag{1'}$$

Vzorci (1) a (1') jsou určeny všechny pravděpodobnosti vztahující se k x a y, je-li dána pravděpodobnost $P(x_i, y_k)$ jako funkce indexů i a k. Součty dle i v předešlých vzorcích i v následujících vztahují se ke všem možným hodnotám x_i, součty dle k ke všem možným hodnotám y_k.

Zaveďme do počtu střední hodnoty veličin x a y:

$$E(x) = \sum_i P(x_i) \cdot x_i, \quad E(y) = \sum_k P(y_k) \cdot y_k, \tag{2}$$

střední hodnoty jejich čtverců:

$$E(x^2) = \sum_i P(x_i) \cdot x_i^2, \quad E(y^2) = \sum_k P(y_k) \cdot y_k^2 \tag{3}$$

a normalisované proměnné ξ, η podle rovnice (4), odst. 53:

$$\xi = \frac{x - E(x)}{\sqrt{E(x^2) - [E(x)]^2}}, \quad \eta = \frac{y - E(y)}{\sqrt{E(y^2) - [E(y)]^2}}. \tag{4}$$

Koeficient korelace R mezi veličinami x a y je roven střední hodnotě součinu normalisovaných veličin ξ, η, tedy

$$R = E(\xi \cdot \eta). \tag{5}$$

Dosadíme-li sem podle (4), bude

$$R = \frac{E\{[x - E(x)] \cdot [y - E(y)]\}}{\sqrt{E(x^2) - [E(x)]^2} \cdot \sqrt{E(y^2) - [E(y)]^2}}; \tag{6}$$

poněvadž

$$E\{[x - E(x)][y - E(y)]\} = E(x \cdot y) - E(x) \cdot E(y),$$

můžeme psáti místo (6) též

$$R = \frac{E(x \cdot y) - E(x) \cdot E(y)}{\sqrt{\{E(x^2) - [E(x)]^2\}\{E(y^2) - [E(y)]^2\}}}. \tag{7}$$

Vzhledem k hořejší definici pravděpodobnosti $P(x_i, y_k)$ je

$$E(x \cdot y) = \sum_i \sum_k P(x_i, y_k)\, x_i y_k; \tag{8}$$

ostatní veličiny $E(x)$, $E(y)$, $E(x^2)$, $E(y^2)$ vyskytující se v rovnici (7) jsou určeny vzorci (2) a (3).

b) *Absolutní hodnota* $|R|$ *koeficientu korelace není nikdy větší než* 1. Abychom to dokázali, utvořme střední hodnotu výrazu $(\xi - \lambda\eta)^2$, kde λ je libovolné reálné číslo a kde ξ a η jsou určeny rovnicemi (4). Vychází

$$E(\xi - \lambda\eta)^2 = E(\xi^2) - 2\lambda E(\xi \,.\, \eta) + \lambda^2 E(\eta^2) \geqq 0.$$

Poněvadž tato nerovnost platí pro každou hodnotu veličiny λ, musí míti mnohočlen druhého stupně vzhledem k λ záporný diskriminant; je tedy

$$[E(\xi \,.\, \eta)]^2 \leqq E(\xi^2) \,.\, E(\eta^2).$$

Vzhledem k rovnici (5) a vzhledem k vlastnostem normalisovaných proměnných ξ a η (viz druhou rovnici (1), odst. 53) je

$$E(\xi\eta) = R, \; E(\xi^2) = E(\eta^2) = 1$$

a tedy

$$R^2 \leqq 1, \; -1 \leqq R \leqq +1.$$

c) Uveďme tři příklady:

První příklad. Mezi veličinami x a y je vztah

$$y = x + \varkappa,$$

kde α je konstanta. Pak je

$$y - E(y) = x + \alpha - E(x) - \alpha = x - E(x)$$

a tedy — viz rovnici (3), odst. 53 —

$$E(y^2) - [E(y)]^2 = E[y - E(y)]^2 = E(x^2) - [E(x)]^2 = \\ = E[x - E(x)]^2. \qquad (9)$$

Koeficient korelace je zde podle (6)

$$R = \frac{E[x - E(x)]^2}{E[x - E(x)]^2} = +1.$$

Kdyby α nebyla konstanta, nýbrž nějaká veličina závislá na náhodě, avšak velmi malá, byl by koeficient R blízký jedné. Dvě veličiny x, y, které se mění přibližně stejně jedna jako druhá (takže rozdíl mezi nimi je malý), mají koeficient korelace blízký kladné jednotce.

Druhý příklad. Je-li mezi x a y vztah

$$y = -x + \alpha,$$

kde α je konstanta, je

$$y - E(y) = -x + \alpha + E(x) - \alpha = -[x - E(x)];$$

rovnice (9) zůstávají v platnosti. Dosadíme-li do (6), vychází

$$R = \frac{-E[x - E(x)]^2}{E[x - E(x)]^2} = -1.$$

Záporný koeficient korelace se vyskytuje v případech, kdy jedna z veličin x, y se zmenšuje, zvětšuje-li se druhá.

Třetí příklad. V osudí jsou koule tří barev; budiž p_i pravděpodobnost vytáhnouti kouli i-té barvy ($i = 1, 2, 3$),

$$p_1 + p_2 + p_3 = 1. \tag{10}$$

Vykonáme n tahů kladouce po každém tahu kouli zpět do osudí. Budiž x počet vytažených koulí prvé barvy a y počet vytažených koulí druhé barvy. Abychom určili obecně koeficient korelace R mezi x a y (x a y mohou nabývati hodnot 0, 1, 2, 3, ..., n), přiřaďme i-tému tahu veličiny u_i a v_i tak, že

$u_i = 1$, vytáhneme-li kouli první barvy v i-tém tahu,

$u_i = 0$, vytáhneme-li kouli druhé nebo třetí barvy v i-tém tahu,

$v_i = 1$, vytáhneme-li kouli druhé barvy v i-tém tahu,

$v_i = 0$, vytáhneme-li kouli první nebo třetí barvy v i-tém tahu.

Pak bude

$$x = u_1 + u_2 + \ldots + u_n, \quad y = v_1 + v_2 + \ldots + v_n.$$

Tahu koule první barvy odpovídají: pravděpodobnost p_1 a hodnoty $u_i = 1$, $v_i = 0$.

Tahu koule druhé barvy odpovídají: pravděpodobnost p_2 a hodnoty $u_i = 0$, $v_i = 1$.

Tahu koule třetí barvy odpovídají: pravděpodobnost p_3 a hodnoty $u_i = 0$, $v_i = 0$.

Případ $u_i = 1$ a $v_i = 1$ není možný. Na základě těchto dat vypočítáme střední hodnoty veličin u_i a v_i pro i-tý tah:

$$E(u_i) = p_1, \quad E(v_i) = p_2, \quad \text{pro } i = 1, 2, \ldots, n.$$

Poněvadž pokusy jsou nezávislé jeden na druhém, je pro $i \neq k$

$$E[(u_i - p_1)(u_k - p_1)] = E(u_i - p_1) \,.\, E(u_k - p_1) = 0. \quad (11)$$

Dále je $E(x) = np_1$, $E(y) = np_2$, $E[x - E(x)] = 0$,

$$E[y - E(y)] = 0,$$

$$E(x^2) - [E(x)]^2 = E[x - E(x)]^2 = E[x - np_1]^2 =$$
$$= E[u_1 - p_1 + u_2 - p_1 + \ldots + u_n - p_1]^2 = n\,E[u_i - p_1]^2 =$$
$$= n[(1 - p_1^2)\,p_1 + p_1^2(1 - p_1)] = np_1(1 - p_1). \quad (12)$$

Podobně se odůvodní, že

$$E(y^2) - [E(y)]^2 = E[y - E(y)]^2 = n\,E[v_i - p_2]^2 =$$
$$= np_2(1 - p_2). \quad (12a)$$

Pokud je $i \neq k$, je

$$E[(u_i - p_1)(v_k - p_2)] = E(u_i - p_1) \,.\, E(v_k - p_2) = 0; \quad (13)$$

střední hodnota součinu $(u_i - p_1)(v_i - p_2)$ rovná se součtu tří členů, které odpovídají po řadě třem shora uvedeným případům:

$$u_i = 1,\ v_i = 0;\ u_i = 0,\ v_i = 1;\ u_i = 0,\ v_i = 0$$

s příslušnými pravděpodobnostmi p_1, p_2, p_3. Je tedy vzhledem k (10)

$$E[(u_i - p_1)(v_i - p_2)] = -(1 - p_1)\,p_2p_1 - p_1(1 - p_2)\,p_2 +$$
$$+ p_1p_2(1 - p_1 - p_2) = -p_1p_2. \quad (14)$$

Z rovnic (11), (12), (12a) a (13) následuje, že

$$E\{[x - E(x)][y - E(y)]\} =$$
$$= E\{[u_1 - p_1 + u_2 - p_1 + \ldots + u_n - p_1] \,.$$
$$\,.\, [v_1 - p_2 + v_2 - p_2 + \ldots + v_n - p_2]\} = \quad (15)$$
$$= -np_1p_2.$$

Dosaďme do (6) příslušné výrazy podle (12), (12a) a (15); vychází

$$R = -\frac{p_1 p_2}{\sqrt{p_1(1-p_1)\,p_2(1-p_2)}} = -\sqrt{\frac{p_1 p_2}{(p_1+p_3)(p_2+p_3)}}.$$

Kdyby nebylo koulí třetí barvy, bylo by $p_3 = 0$, $x + y = n$; y by bylo zcela určitou funkcí proměnné x a koeficient korelace R by byl roven -1. Je-li p_3 veličina malá proti p_1 a p_2 (je-li tedy koulí třetí barvy velmi málo), platí rovnice $x + y = n$ jen přibližně, koeficient R se liší málo od -1. Jsou-li naopak p_1 a p_2 čísla malá proti p_3 (v osudí je jen málo koulí první a druhé barvy, převládají koule třetí barvy), je R přibližně rovno nule; mezi x a y není určitého vztahu, tahy koulí první a druhé barvy jsou vzácné a nezávislé jedny na druhých.

55. Empirické stanovení koeficientu korelace. V odst. 54a jsme předpokládali, že jsou známy pravděpodobnosti $P(x_i, y_k)$, ze kterých jsme odvodili další pravděpodobnosti pro výskyt znaků x_i resp. y_k a pak koeficient korelace R. V empirických problémech nejsou však dány přímo pravděpodobnosti $P(x_i, y_k)$, nýbrž statistická data o výskytu znaků, které sestavujeme v t. zv. *korelační tabulku*. Budiž n_{ik} počet případů, kdy první proměnná x má hodnotu x_i a kdy zároveň druhá veličina y má hodnotu y_k; $i = 1, 2, \ldots$, $k = 1, 2, \ldots$ Veličiny n_{ik} jsou dány; považujeme-li i za index řádku a k za index sloupce, napíšeme n_{ik} do korelační tabulky na místo, kde se i-tý řádek protíná s k-tým sloupcem. Empirické hodnoty pravděpodobností čárkovanými označíme písmeny. Empirická pravděpodobnost $P'(x_i, y_k)$, že $x = x_i$ a že $y = y_k$, je dána vzorcem

$$P'(x_i, y_k) = \frac{n_{ik}}{\sum_i \sum_k n_{ik}}.$$

Dále odvodíme z korelační tabulky, užívajíce označení obdobného tomu, které jsme zavedli v odst. 54a, tyto empirické pravděpodobnosti.

Pravděpodobnost $P'(x_i)$, že $x = x_i$, je rovna

$$\frac{\sum_k n_{ik}}{\sum_i \sum_k n_{ik}}.$$

Pravděpodobnost $P'(y_k)$, že $y = y_k$, je rovna

$$\frac{\sum_i n_{ik}}{\sum_i \sum_k n_{ik}}.$$

Pravděpodobnost $P_{y_k}'(x_i)$, že $x = x_i$, je-li dáno, že $y = y_k$, je rovna

$$\frac{n_{ik}}{\sum_i n_{ik}}.$$

Pravděpodobnost $P'_{x_i}(y_k)$, že $y = y_k$, je-li dáno, že $x = x_i$, je rovna

$$\frac{n_{ik}}{\sum_k n_{ik}}.$$

Příslušné empirické střední hodnoty součinu xy veličin x a y a jejich čtverců jsou

$$E'(xy) = \frac{\sum_i \sum_k n_{ik} x_i y_k}{\sum_i \sum_k n_{ik}}, \quad E'(x) = \frac{\sum_i \sum_k n_{ik} x_i}{\sum_i \sum_k n_{ik}}, \quad E'(y) = \frac{\sum_i \sum_k n_{ik} y_k}{\sum_i \sum_k n_{ik}},$$

$$E'(x^2) = \frac{\sum_i \sum_k n_{ik} x_i^2}{\sum_i \sum_k n_{ik}}, \quad E'(y^2) = \frac{\sum_i \sum_k n_{ik} y_k^2}{\sum_i \sum_k n_{ik}}.$$

Empirický koeficient korelace R' se vypočte z těchto empirických hodnot právě tak, jako jsme vypočetli theoretický koeficient korelace R na základě theoretických středních hodnot podle vzorce (6), odst. 54. Je tedy

$$R' = \frac{E'(xy) - E'(x) \,.\, E'(y)}{\sqrt{\{E'(x^2) - [E'(x)]^2\}\{E'(y^2) - [E'(y)]^2\}}} \qquad (1)$$

nebo, dosadíme-li podle předešlých vzorců za $E'(xy)$, $E'(x)$, $E'(y)$, ...,

$$R' = \frac{N \,.\, \sum_i \sum_k n_{ik} x_i y_k - (\sum_i \sum_k n_{ik} x_i)(\sum_i \sum_k n_{ik} y_k)}{\sqrt{[N \,.\, \sum_i \sum_k n_{ik} x_i^2 - (\sum_i \sum_k n_{ik} x_i)^2][N \,.\, \sum_i \sum_k n_{ik} y_k^2 - (\sum_i \sum_k n_{ik} y_k)]}}, \qquad (2)$$

kde jsme položili pro stručnost

$$N = \sum_i \sum_k n_{ik}.$$

Jsou-li tedy dány: korelační tabulka, t. j. hodnoty n_{ik}, ($i = 1, 2, \ldots$, $k = 1, 2, \ldots$), a hodnoty $x_1, x_2, \ldots, y_1, y_2, \ldots$, vypočte se empirická hodnota koeficientu korelace podle (2).*)

Poznámka. V odst. 54c jsme viděli, že zvláštní druhy závislosti mezi x a y odpovídají různým hodnotám koeficientu R. Je-li naopak dána empirická korelační tabulka, určíme z ní podle (2) empirický koeficient korelace R'. Ptáme se pak, jaký závěr možno učiniti z číselné hodnoty R' na povahu vztahu mezi x a y. Odpověď zní, že jediná číselná hodnota R' nestačí k tomu, aby charakterisovala závislost mezi x a y po všech stránkách.

56. Kvalitativní koeficient korelace. Podle definice (7) odst. 54 je koeficient korelace R funkcí veličin $x_1, x_2, \ldots$ a $y_1, y_2, \ldots$, při čemž pravděpodobnosti $P(x_i, y_k)$, $P(x_i)$ a $P(y_k)$ mají úlohu koeficientů. Naskytuje se otázka, může-li R býti veli-

*) Číselné příklady korelačních tabulek uvádí *S. Kohn* ve spise Základy teorie statistické metody (Praha 1929) na str. 297. Viz též *J. Kaucký*: Úvod do počtu pravděpodobnosti a teorie statistiky, str. 49, Praha 1934. *J. Kaucký-J. Novák-Vl. List*: Užití korelačního počtu, Praha 1948.

čina nezávislá na proměnných $x_1, x_2, \ldots, y_1, y_2, \ldots$ Vyšetříme tuto otázku pro případ, že veličina x může nabýti jen dvou různých hodnot x_1, x_2 a veličina y také jen dvou hodnot y_1, y_2. Pro jednoduchost volíme o něco stručnější označení pravděpodobností; zavedeme na místo znaků definovaných v odst. 54a,

$$p_{ik} = P(x_i, y_k),\ i = 1, 2,\ k = 1, 2;$$
$$p_i = P(x_i),\ i = 1, 2;\ p_k' = P(y_k),\ k = 1, 2.$$

Pak bude

$$E(xy) = p_{11}x_1y_1 + p_{12}x_1y_2 + p_{21}x_2y_1 + p_{22}x_2y_2,$$
$$E(x) = p_1x_1 + p_2x_2,\ E(y) = p_1'y_1 + p_2'y_2,$$
$$E(x^2) = p_1x_1^2 + p_2x_2^2,\ E(y^2) = p_1'y_1^2 + p_2'y_2^2,$$

při čemž

$$\left.\begin{aligned} p_1 &= p_{11} + p_{12},\ p_2 = p_{21} + p_{22},\ p_1 + p_2 = 1,\\ p_1' &= p_{11} + p_{21},\ p_2' = p_{12} + p_{22},\ p_1' + p_2' = 1. \end{aligned}\right\} \quad (1)$$

Koeficient korelace se pak vyjádří podle (7), odst. 54 takto:

$$R = \frac{\sum_{i=1}^{2}\sum_{k=1}^{2} p_{ik}x_ix_k - (\sum_{i=1}^{2} p_ix_i)(\sum_{k=1}^{2} p_k'y_k)}{\sqrt{[\sum_{i=1}^{2} p_ix_i^2 - (\sum_{k=1}^{2} p_kx_k)^2][\sum_{i=1}^{2} p_i'y_i^2 - (\sum_{k=1}^{2} p_k'y_k)^2]}}. \quad (2)$$

V čitateli je mnohočlen druhého stupně vzhledem k proměnným x_i a y_k. Ve jmenovateli je odmocnina ze součinu dvou mnohočlenů proměnných x_i resp. y_k. Proto nemůže býti R obecně veličinou nezávislou na proměnných x_1, x_2, y_1 a y_2. Případ nezávislosti může nastati jen tehdy, redukuje-li se výraz pod odmocninou jmenovatele na druhou mocninu mnohočlenu stojícího v čitateli násobenou nějakou konstantou. Aby tento případ nastal, je předně nutno položiti

$$y_1 = x_1,\ y_2 = x_2,\ p_1' = p_1,\ p_2' = p_2;$$

za tohoto předpokladu přejde (2) ve formuli

$$R = \frac{(p_{11} - p_1^2)\,x_1^2 + (p_{12} + p_{21} - 2p_1p_2)\,x_1x_2 + (p_{22} - p_2^2)\,x_2^2}{(p_1 - p_1^2)\,x_1^2 - 2p_1p_2x_1x_2 + (p_2 - p_2^2)\,x_2^2}. \tag{3}$$

R je tedy funkcí poměru $x_2 : x_1$. Aby R nezávisel vůbec na hodnotě tohoto poměru, je nutno a stačí položiti

$$\frac{p_{11} - p_1^2}{p_1 - p_1^2} = \frac{p_{12} + p_{21} - 2p_1p_2}{-2p_1p_2} = \frac{p_{22} - p_2^2}{p_2 - p_2^2} = R. \tag{4}$$

Jsou-li splněny tyto rovnice, mají mnohočleny druhého stupně vyskytující se v čitateli a ve jmenovateli výrazu (3) úměrné koeficienty; poměr souhlasných koeficientů je roven konstantě R, která nezávisí na hodnotách x_1, x_2.

Rovnicím (4) vyhoví se takto: Volíme nejprve p_1, p_2 a R tak, aby

$$p_1 \geqq 0,\ p_2 \geqq 0,\ p_1 + p_2 = 1,\ -1 \leqq R \leqq +1; \tag{5}$$

pak ustanovíme p_{ik} rovnicemi:

$$\begin{aligned} p_{11} &= p_1^2 + R(p_1 - p_1^2), \\ p_{12} &= p_{21} = p_1p_2(1 - R), \\ p_{22} &= p_2^2 + R(p_2 - p_2^2). \end{aligned} \tag{6}$$

Jsou-li splněny rovnice (5) a (6), je vyhověno i prvním dvěma rovnicím (1). Neboť

$$\begin{aligned} p_{11} + p_{12} &= p_1^2 + p_1p_2 + R(p_1 - p_1^2 - p_1p_2) = \\ &= p_1 + R[p_1 - p_1^2 - p_1(1 - p_1)] = p_1 \end{aligned}$$

a podobně se odůvodní, že

$$p_{21} + p_{22} = p_2.$$

Výsledek shrneme takto:

Budiž R koeficient korelace mezi dvěma veličinami x, y, z nichž první nabývá hodnot x_1, x_2 a druhá y_1, y_2. Platí-li v označení zavedeném v tomto odstavci

$$y_1 = x_1,\ y_2 = x_2,\ p_1' = p_1,\ p_2' = p_2$$

a jsou-li splněny podmínky (5) *a* (6), *je R veličina nezávislá na* x_1 *a na* x_2.

Veličina R se pak nazývá *kvalitativní koeficient korelace.* V případě, že x a y nabývají více hodnot než dvou, vyskytují se také kvalitativní koeficienty korelace.

Poznámka. Zjednodušení úlohy, jehož bylo docíleno druhou rovnicí (6), totiž předpoklad, že $p_{12} = p_{21}$, odpovídá případu, kdy běží o koeficient korelace mezi veličinami $x^{(m)}$ a $x^{(n)}$ přiřaděnými dvěma pokusům v případě stacionárního řetězu (viz odst. 82c).