Jak vytváří statistika obrazy světa a života. II. díl

Část II. Korelace [odst. 5,1-5,3,6,1-6,5,7,1-7,2, 8,1-8,4,9,1-9,5,10]

In: Jaroslav Janko (author): Jak vytváří statistika obrazy světa a života. II. díl. (Czech). Praha: Jednota českých matematiků a fysiků, 1944. pp. 83–153.

ČÁST II.

KORELACE.

Dosud jsme se věnovali vlastnostem rozdělení četnosti jednoho znaku, které jsme odvodili z posloupnosti hodnot pozorovaných na prvcích souboru. Jestliže uvažujeme na př. soubor listů natrhaných s určitého stromu a měříme jejich délku, dostaneme rozdělení četností hodnot znaku „délka listu tohoto stromu". Takové rozdělení četností jsme nazvali jednorozměrným. Nyní přecházíme k vícerozměrnému rozdělení četností, především k dvojrozměrnému. Slovo korelace znamená totiž vzájemný vztah a teorie korelace, jíž se budeme zabývati studuje vzájemný vztah statistických řad, což je úkol nový, který se při rozdělení četností podle jednoho znaku nevyskytoval a je jedním z nejdůležitějších problémů statistiky.

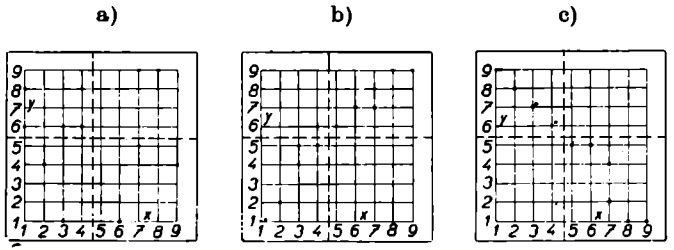
(5,1) Pojem korelace. Představme si, že měříme na prvku určitého souboru hodnoty dvou znaků x a y, takže pro všechny prvky souboru rozsahu r dostaneme r párů hodnot $x_1, y_1; \ldots; x_r, y_r$. Pozorovali jsme na př. páry hodnot uvedené v tab. 10, kde jsou tři různé soubory, každý rozsahu $r = 13$.

Sledujeme-li průběh hodnot znaků v dvojicích jednotlivých souborů, vidíme, že v souboru č. 3 jsou sdruženy nízké hodnoty x s vysokými hodnotami y, kdežto v souboru č. 2 jsou sdruženy vysoké hodnoty x s vysokými hodnotami y. V obou případech je tu patrný vztah mezi oběma znaky, ale jeden je obráceným vztahem druhého. V souboru č. 1 nepoznáváme zřejmého vztahu mezi oběma proměnnými x, y.

Obraz těchto poměrů dostaneme na tečkovém diagramu (obr. 7), v němž značí hodnoty x úsečky bodů a hodnoty y jejich pořadnice. Je tedy každý pár hodnot x, y znázorněn tečkou a vztah mezi hodnotami x a y je naznačen všeobecně způsobem rozptýlení teček. Jedná-li se o soubory velkého

Tabulka 10.

Soubor					
č. 1		č. 2		č. 3	
x	y	x	y	x	y
8	9	9	9	1	9
4	8	8	9	1	9
7	5	7	8	2	8
7	9	7	7	3	7
1	6	6	7	3	7
2	4	5	6	4	6
6	1	4	6	4	6
5	3	4	5	5	5
3	1	3	5	6	5
4	6	3	4	7	4
9	4	2	2	7	2
3	6	1	1	8	1
1	8	1	1	9	1
60	70	60	70	60	70



Obr. 7. Tečkové diagramy souborů v tab. 10.

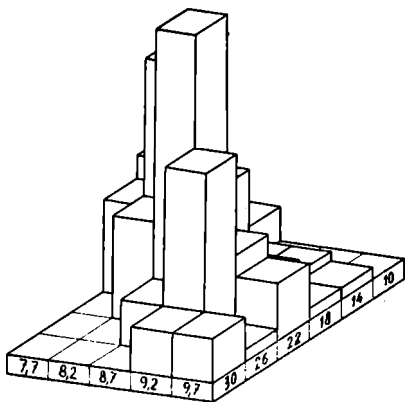
rozsahu, dostáváme roje teček ne nepodobné mléčné dráze. V případech, jakým je náš první soubor, jsou tečky rozhozeny více méně stejnoměrně po celé ploše, takže roj směřuje k tvaru kruhovému, kdežto v případech druhého a třetího souboru se tečky kupí k jedné nebo druhé úhlopříčce.

Vraťme se nyní k souboru listů a ptejme se zda je nějaký vztah mezi délkou a šířkou listu. Kdybychom z něho vzali jeden list, který by byl dlouhý a mohli souditi jen na základě této informace, že je také široký, znamenalo by to, že můžeme předpokládat, že jsou tu nějaké příčiny, pro které jsou délka a šířka listu ve vztahu čili v korelaci. Nemůžeme-li jen podle délky listu odhadnouti jeho šířku, musíme předpokládati, že nejsou ve vztahu. Je patrno, že dva znaky mohou k sobě býti vázány různou silou, čili těsnost vztahu může míti různé stupně, takže chceme těsnost korelace srovnávat a řadit podobně jako jsme dosud seřadili prvky podle velikosti kvantitativního znaku. K tomu cíli sestrojili statistikové zvláštní stupnici, kterou si vysvětlíme. Víme-li, že šířka nějakého listu je přesně polovinou jeho délky nebo že rozdíl mezi délkou a šířkou je vždy 2 cm, řekneme, že délky a šířky listů jsou pevně k sobě vázány, čili jsou v absolutním vzájemném vztahu. Kdekoliv je takový pevný vztah čili funkční vztah mezi dvěma znaky, říkáme, že korelace je úplná (perfektní). Tento pevný vztah je na vrcholu naší stupnice a je označen jednotkou jakožto „koeficientem úplné korelace".

Kdybychom věděli, že délka listu byla jen přibližně dvojnásobkem jeho šířky, tedy někdy trochu více a někdy méně, měli bychom vztah přibližný, který je volnější než dříve uvedený vztah pevný. Koeficient, který jej bude vyznačovati, bude na stupnici někde níže pod jednotkou; k jeho stanovení si zavedeme dále určitou metodu výpočtu. Napřed se ještě zabývejme případem, kdy není vůbec vztahu mezi znaky. Pro znak „šířka listu" si stanovíme průměr celého souboru $\overline{y}$. Pak rozdělíme všechny listy souboru na tři skupiny, takže do první dáme všechny nejdelší, do druhé listy prostřední délky a do třetí všechny krátké listy. Dostaneme-li v každé skupině pro šířku listu týž průměr $\overline{y}$, pak vidíme, že údaj o délce listu nám nepřispěje ničím k odhadu jeho šířky, neboť v takovém případě není vztahu mezi délkou a šířkou. Říkáme, že koeficient korelace je nula.

Z těchto úvah by vyplývalo, že lze každou těsnost vztahu

zařadit na stupnici čísel od 0 do 1, ale nepředstavili jsme si ještě všechny možné případy. Zjistíme-li, že kdykoliv byl nějaký list dlouhý, byl také úzký a kdykoliv byl krátký, byl široký (případ souboru č. 3), máme zase dřívější vztah mezi znaky „dlouhý" a „úzký". Poněvadž však znak „úzký" je vlastně znak „široký" s obráceného hlediska, zahrneme tento případ mezi dřívější, rozšíříme-li stupnici do záporných čísel až k —1.



Obr. 8. Stereogram rozdělení četností v tab. 14.

Jednorozměrné rozdělení četností jsme znázorňovali v rovině histogramem nebo křivkou tak, že jeden rozměr byl vyčerpán stupnicí hodnot znaku a druhý rozměr stupnicí četností jednotlivých hodnot znaku. Při dvojrozměrném rozdělení četností jsou oba rozměry vyčerpány dvojicemi hodnot znaků x a y, takže plocha četnosti může býti znázorněna ve třech rozměrech t. zv. stereogramem (obr. 8).

Je-li rozdělení četností v jednotlivých řádcích i v jednotlivých sloupcích dáno normální křivkou (I, str. 81), je rozdělení četností dvojic hodnot proměnných vyjádřeno normál-

Tabulka 11.

x \ y	Šířka listů → 2,2-	2,4-	2,6-	2,8-	3,0-	3,2-	3,4-	3,6-	3,8-	4,0-	4,2-	Celkem
Délka listů ↓ 4,1—	—	—	1	—	—	1	—	—	—	—	—	2
4,3—	1	—	—	3	—	—	—	—	—	—	—	4
4,5—	1	1	2	2	1	1	—	—	—	—	—	8
4,7—	—	—	2	3	7	5	1	—	—	—	—	18
4,9—	—	1	1	5	9	6	1	—	—	—	—	23
5,1—	—	—	—	7	16	5	2	1	—	—	—	31
5,3—	—	—	—	2	17	8	6	4	—	—	—	37
5,5—	—	—	—	2	6	7	7	1	2	—	1	26
5,7—	—	—	—	1	2	6	5	1	—	—	—	15
5,9—	—	—	—	—	3	5	9	2	1	—	—	20
6,1—	—	—	—	—	—	3	3	1	—	1	—	8
6,3—	—	—	—	—	—	—	2	—	1	—	1	4
6,5—	—	—	—	—	1	—	—	—	1	1	—	3
6,7—	—	—	—	—	—	—	1	—	—	—	—	1
Celkem	2	2	6	25	62	47	37	10	5	2	2	200

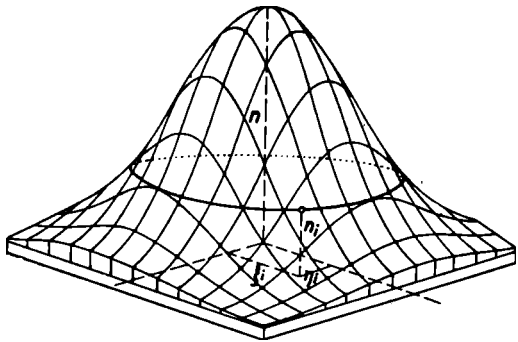
ní plochou dvou proměnných čili normální korelační plochou, která má obdobný význam v teorii dvojrozměrného rozdělení četností jako má normální křivka v teorii rozdělení četností jedné proměnné. Třeba poukázati hlavně na její historický význam, poněvadž s počátku byla teorie korelace budována na předpokladu takového rozdělení. Potom ustoupil tento význam do pozadí, když byly hlavní výsledky odvozeny bez předpokladu o formě rozdělení četností, ale zobecněná forma poskytuje možnost zjednodušeného vyjádření rozptylů charakteristik v teorii náhodného výběru.

Grafické znázornění je provedeno v obr. 9; o teorii této plochy se čtenář blíže doví v [1].

Obtížnost prostorového znázorňování se překonává buď uvedeným již diagramem tečkovým, kde četnost dvojic hodnot znaku je znázorněna hustotou teček na plošce určitého rozměru, nebo dvojrozměrnou tabulkou rozdělení četností,

kde na této plošce je přímo číslicemi zapsána absolutní či relativní četnost dotyčné dvojice hodnot znaků x a y.

Jedná-li se o znak rozpojitý, kde proměnná nabývá jen isolovaných hodnot, na př. jen čísel celých, jako je tomu v tab. 10, pak padnou tečky vždy jen do mřížových bodů (obr. 7) daných hodnotami souřadnic (x, y). Pro znaky spojité jako na př. délka, šířka, musíme (I, str. 29) zavésti třídní intervaly a tečky padnou do pole pravoúhelníka, v němž se překrývají pásy příslušných intervalů (tab. 11).



Obr. 9. Normální plocha korelační.

Do tabulky o dvojím vstupu bychom mohli seřaditi hodnoty x, y kteréhokoliv ze tří uvedených souborů, ale zvolíme si vhodněji rozsáhlejší soubor. Proto jsme si sestavili výsledky pozorování dvou znaků — „délka" a „šířka" — na listech jako prvcích našeho souboru. Čteme-li tuto tabulku 11, vidíme, že na př. ve třetím sloupci je 6 listů šířky 2,6 cm až 2,8 cm (při čemž listy šířky 2,8 cm jsou již ve vedlejším sloupci); z nich je jeden délky mezi 4,1 až 4,3 cm, dva mezi 4,5 až 4,7 cm, dva od 4,7 do 4,9 cm, a jeden mezi 4,9 až 5,1 cm. Podobně na třetím řádku zdola je patrno, že ze čtyř listů délky mezi 6,3 až 6,5 cm jsou dva šířky od 3,4 do 3,6 cm, jeden 3,8 až 4,0 cm a jeden šířky 4,2 až 4,4 cm. Rozdělíme-li

si listy podle délky na tři skupiny: na krátké třeba do 4,9 cm, prostřední do 5,9 cm a dlouhé ostatní, shledáme pozorováním tabulky, že dlouhé listy jsou v celku značně širší než krátké. Přesněji se pak o tom přesvědčíme, jestliže si vypočítáme pro různé délky listů příslušnou průměrnou šířku. Jako třídní znak hodnot znaku x (délka listu) budeme bráti 4,2; 4,4; ... znaku y pak 2,3; 2,5; ...

Tabulka 12.

x_i	$\bar{y}_i$
méně	
než 4,9	2,96
5,0	3,04
5,2	3,13
5,4	3,26
5,6	3,38
5,8	3,34
6,0	3,43
6,2	3,53
nad 6,3	3,73

Tabulka 13.

y_k	$\bar{x}_k$
méně	
než 2,8	4,66
2,9	5,04
3,1	5,29
3,3	5,42
3,5	5,76
nad 3,6	5,86

Máme tedy pro určitou hodnotu x_i jednorozměrné rozdělení četností hodnot y na dotyčném řádku a průměry těchto řádků $\bar{y}_i$ jsou sestaveny v druhém sloupci tab. 12. Z výsledků je viděti, že zatím co délka listů x vzrostla přibližně se 4,4 do 6,6 cm, tedy o 2,2 cm, vzrostla souběžně příslušná průměrná šířka s 2,96 cm na 3,73 cm, čili o 0,77 cm.

Podobně dostaneme tab. 13, v níž jsou sestaveny k jednotlivým hodnotám šířky y_k příslušné průměrné délky $\bar{x}_k$. Je tedy ve druhém sloupci vždy uvedena průměrná délka všech listů, které mají šířku uvedenou vedle v prvním sloupci. Tak odpovídá vzrůst průměrných délek o 1,20 cm vzrůstu asi o 1,4 cm skutečných šířek.

Tyto dva výsledky nám podávají vztah délky listů k průměrné šířce a šířky listu k průměrné délce. Je-li při úplné

korelaci na př. šířka listu polovinou délky, je délka listu dvojnásobkem šířky jednotlivě i v průměru, takže vázanost šířky k délce a rovněž délky k šířce je úplná. Bylo by tedy nasnadě očekávati, že bude vhodnou mírou korelace poměr $\frac{0,77}{2,2}$ a rovněž poměr $\frac{1,20}{1,4}$. Je jasno, že těsnost vztahu délky k šířce je táž jako vztahu šířky k délce; uvedené poměry se však od sebe velmi liší, takže je zřejmo, že jsou k tomuto účelu nevhodné.

Snažíme-li se opravit tato poměrná čísla, všimneme si ihned, že jsme zvolili v tabulce třídní intervaly 0,2 cm pro délky listů i pro šířky a ptáme se, zda by nebylo správnější vzíti pro délky 0,2 cm a pro šířky 0,1 cm. Je nám totiž zřejmo, že by se skutečná míra korelace nezměnila, kdybychom na př. měřili délky listů v palcích a šířky v cm, ale naše zlomky by se jistě změnily. Vidíme tedy, že délky třídních intervalů nejsou vhodné k našemu účelu. Správnější se ukazuje volba směrodatné odchylky za jednotku měření, takže pak uvažujeme směrodatné proměnné.

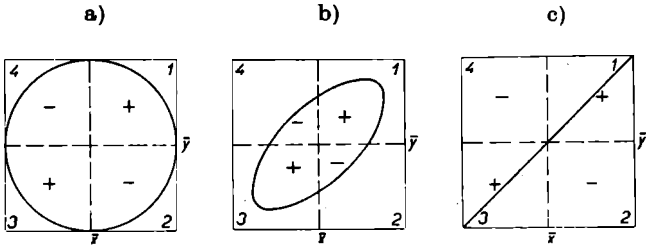
Výsledek tohoto postupu si osvětlíme na našem příkladu. Vypočítáme-li směrodatnou odchylku znaku x, čili délek listů bez ohledu na jejich šířku, použijeme četností krajního (marginálního) sloupce a dostaneme (podle I, str. 22) $\sigma_x = 0,50$. Podobně z marginálního řádku je $\sigma_y = 0,31$. Skutečně tedy budeme měřit délku přibližně dvojnásobnou jednotkou než šířku a provedeme-li měření, dostáváme pro přibližné variační obory

a) délky	$2,2 : 0,50 = 4,40$	takže poměr	$\frac{4,40}{2,48} = 1,77$
prům. šířky	$0,77 : 0,31 = 2,48$		
b) šířky	$1,40 : 0,31 = 4,52$	a poměr	$\frac{4,52}{2,40} = 1,88$
prům. délky	$1,20 : 0,50 = 2,40$		

Tyto dva zlomky jsou nyní prakticky stejné a dávají tedy vhodnou míru těsnosti vztahu mezi délkami a šířkami listů.

Této přibližné metody a odhadu variačních šířek s ohledem na velmi slabě obsazené kraje jsme použili jen proto, že se osvědčila k výkladu smyslu funkce užité k měření korelace. Skutečné měření korelace se provádí jinými metodami, z nichž hlavní nyní objasníme.

(5,2) Měření korelace. K podání soustředěné informace o určitém souboru, obsažené v jednorozměrném rozdělení četností, užíváme systému charakteristik, v němž nejdůležitějšími jsou průměr a směrodatná odchylka. Podobný systém charakteristik zavádíme pro dvojrozměrné rozdělení četností.



Obr. 10. Součiny odchylek $\xi\eta$ v jednotlivých kvadrantech při rostoucím stupni kladné vázanosti.

Vezměme tedy v úvahu soubor č. 1 z tab. 10. Rozdělení četností hodnot znaku x má svůj průměr $\bar{x} = 4{,}6$ a směrodatnou odchylku $\sigma_x = 2{,}5$. Rozdělení četností hodnot znaku y pak má průměr $\bar{y} = 5{,}4$ a směrodatnou odchylku $\sigma_y = 2{,}6$. K nim přistupuje dále nová charakteristika, která bude zahrnovati současně hodnoty obou znaků resp. hodnoty jejich odchylek od příslušného průměru. Rozdělme celé pole prvního tečkového roje v obr. 7 osami pravoúhlých souřadnic, jež mají počátek v bodě daném průměry ($\bar{x}$, $\bar{y}$) na čtyři kvadranty (viz obr. 10a). V prvním kvadrantu jsou kladné odchylky ($x - \bar{x}$) od průměru $\bar{x}$ a kladné odchylky $y - \bar{y}$ od průměru $\bar{y}$. V druhém kvadrantu jsou kladné odchylky ($x - \bar{x}$) a záporné odchylky ($y - \bar{y}$), ve třetím jsou záporné odchylky

$(x - \bar{x})$ i záporné odchylky $(y - \bar{y})$ a ve čtvrtém kvadrantu jsou záporné odchylky $(x - \bar{x})$ s kladnými odchylkami $(y - \bar{y})$. Každý bod roviny je v těchto souřadnicích vyznačen párem odchylek $(x_i - \bar{x}, y_k - \bar{y})$.

Utvořme nyní součin těchto odchylek $(x_i - \bar{x})(y_k - \bar{y})$. Pak vidíme, že v prvním a třetím kvadrantu mají součiny znaménko kladné, kdežto v druhém a čtvrtém kvadrantu mají znaménko záporné. Jsou-li body rozděleny stejnoměrně ve všech kvadrantech, pak součet součinů odchylek $\Sigma(x_i - \bar{x})(y_k - \bar{y})$ kladných se rovná součtu součinů odchylek záporných, takže v celkovém součtu všech se součiny vzájemně zruší a výsledek bude roven nule. Takovému roji říkáme kruhový a představuje nulovou korelaci. Je-li rozdělení bodů jen přibližně stejnoměrné, tedy roj jen přibližně kruhový, jako je tomu v obr. 7a, bude celkový součet součinů malé číslo, které nasvědčuje tomu, že není mezi oběma znaky vztahu. Druhý krajní případ nastává, je-li celý roj teček na úhlopříčce (obr. 10c), takže všechny součiny jsou kladné a součet může dosáhnouti největší hodnoty. Mezi oběma uvedenými případy pak je ten, kde kladné součiny jsou větší než záporné, nebo aspoň jejich součet, takže máme kladnou korelaci ovšem neúplnou; je znázorněna obr. 10b.

Souhlas skutečnosti s těmito úvahami můžeme ukázati na souborech tab. 10, vypočítáme-li pro každý z nich

$$\Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - \bar{x}\,\Sigma y - \bar{y}\,\Sigma x + r\bar{x}\bar{y} =$$
$$= \Sigma xy - r\bar{x}\bar{y} = \Sigma xy - \frac{\Sigma x\,\Sigma y}{r}.$$

Pro zjednodušení a přehlednost výrazů nebudeme v dalším, pokud toho nebude zvlášť třeba, vyznačovati indexy jednotlivých hodnot proměnných a součtové meze, takže symbol Σ bude značit součet všech v souboru se vyskytujících hodnot dotyčné proměnné, která je za součtovým znaménkem nebo součet všech v souboru se vyskytujících dvojic, jsou-li za součtovým znaménkem symboly dvou proměnných.

Tak dostaneme pro

	Σxy	$\frac{1}{r}\Sigma x\,\Sigma y$	$\Sigma(x-\bar{x})(y-\bar{y})$
soubor č. 1	326	323	3
soubor č. 2	407	323	84
soubor č. 3	238	323	—85.

Vidíme skutečně, že součet součinů odchylek je u prvního souboru, kde jsme nemohli poznati, zda je mezi oběma znaky nějaký vztah, velmi malý, kdežto u souborů č. 2 a č. 3, kde je vztah zřejmý, je tento součet co do absolutní hodnoty velký a v případě souboru č. 3, kde jsou sdruženy malé hodnoty znaku x s velkými hodnotami znaku y má znaménko minus, tedy obrácené než v případě souboru č. 2. Je tedy tento součin mírou korelace, ale poněvadž při témž stupni těsnosti roste s variabilitou hodnot znaků, ukázali jsme si již, že je třeba měřiti každou odchylku od průměru ve směrodatné odchylce jako jednotce, která je nejvhodnější mírou variability. Tak dostaneme konečně koeficient korelace r_{xy} v jeho nejzákladnějším tvaru

$$r_{xy} = \frac{1}{r}\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right).$$

Pro naše soubory najdeme tedy hodnoty

$${}_1r_{xy} = 0{,}03, \quad {}_2r_{xy} = 0{,}97, \quad {}_3r_{xy} = -0{,}98.$$

Provedeme ještě jednoduchý důkaz o extrémních hodnotách, t. j., že koeficient korelace má své maximum $+1$ a minimum -1.

Předpokládejme, že je konstantní (úplný) kladný vztah mezi znaky x a y, že tedy platí $y = cx$, kde c je konstanta. Pak budou odchylky

$$y - \bar{y} = cx - c\bar{x} = c(x - \bar{x}),$$
$$\Sigma(y - \bar{y}) = c\Sigma(x - \bar{x}),$$

a součet součinů

$$\Sigma(x - \bar{x})(y - \bar{y}) = c\Sigma(x - \bar{x})^2,$$

dále

$$\Sigma(y - \bar{y})^2 = c^2\Sigma(x - \bar{x})^2 \text{ a tedy } \sigma_y = c\sigma_x,$$

koeficient korelace

$$\frac{\frac{1}{r}\Sigma(x - \bar{x})(y - \bar{y})}{\sigma_x\sigma_y} = \frac{c\frac{1}{r}\Sigma(x - \bar{x})^2}{c\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = +1.$$

Zcela obdobně vyplývá hodnota —1, je-li mezi x a y konstantní negativní vztah, tedy $y = -cx$, což si čtenář laskavě sám provede.

(5,3) Lineární regrese. Uvažujme nyní vztah mezi proměnnými x a y, jak se jeví v souboru č. 2, tab. 10; z grafického znázornění (obr. 7b) se nám jeví přibližně lineárním. Pokusíme se jej tedy vyjádřit přímkou, která se bodům (x_i, y_k) přimyká tak, že součet čtverců odchylek od ní, rovnoběžných s osou y, je nejmenší. Budeme ji nazývati „přímka odhadu", neboť nám pomáhá odhadovat hodnoty jedné proměnné na základě znalosti hodnot druhé proměnné. Rovnici této přímky píšeme ve tvaru $y = ax + b$. Kdyby všechny body ležely na této přímce, splňovalo by všech třináct dvojic hodnot x_i, y_k tuto rovnici. Dostali bychom tedy třináct rovnic

$$\begin{aligned}
9 &= 9a + b & 6 &= 4a + b\\
9 &= 8a + b & 5 &= 4a + b\\
8 &= 7a + b & 5 &= 3a + b\\
7 &= 7a + b & 4 &= 3a + b\\
7 &= 6a + b & 2 &= 2a + b\\
6 &= 5a + b & 1 &= 1a + b\\
 & & 1 &= 1a + b.
\end{aligned}$$

Abychom našli metodou nejmenšího součtu čtverců hodnoty a, b tak, aby přímka vyhovovala uvedené podmínce, na-

jdeme t. zv. normální rovnice. První dostaneme sečtením všech rovnic, čili

$$\Sigma y_i = a\Sigma x_i + rb, \tag{62}$$

a druhou dostaneme, jestliže každou z rovnic násobíme příslušnými x_i a všechny opět sečteme, takže

$$\Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i, \tag{63}$$

a v našem případě budou tedy normální rovnice

$$\begin{aligned} 70 &= 60a + 13b \\ 407 &= 360a + 60b, \end{aligned}$$

z nichž plynou hodnoty

$$a = 1{,}01, \quad b = 0{,}72,$$

takže rovnice přímky bude

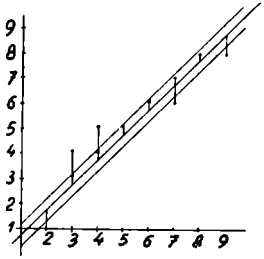
$$y = 1{,}01x + 0{,}72.$$

Z ní vypočítáme ke každému x_i, které bylo pozorováno na prvcích uvažovaného souboru příslušné y_i. Rozdíl pak mezi skutečně pozorovanou hodnotou znaku y a takto vypočítanou se nazývá odchylkou residuální e_i. Bude tedy

$$\begin{aligned} e_1 &= 9 - 1{,}01 \times 9 - 0{,}72 = -0{,}81 \\ e_2 &= 9 - 1{,}01 \times 8 - 0{,}72 = +0{,}20 \\ e_3 &= 8 - 1{,}01 \times 7 - 0{,}72 = +0{,}21 \\ e_4 &= 7 - 1{,}01 \times 7 - 0{,}72 = -0{,}79 \\ e_5 &= 7 - 1{,}01 \times 6 - 0{,}72 = +0{,}22 \\ e_6 &= 6 - 1{,}01 \times 5 - 0{,}72 = +0{,}23 \\ e_7 &= 6 - 1{,}01 \times 4 - 0{,}72 = +1{,}24 \\ e_8 &= 5 - 1{,}01 \times 4 - 0{,}72 = +0{,}24 \\ e_9 &= 5 - 1{,}01 \times 3 - 0{,}72 = +1{,}25 \\ e_{10} &= 4 - 1{,}01 \times 3 - 0{,}72 = +0{,}25 \\ e_{11} &= 2 - 1{,}01 \times 2 - 0{,}72 = -0{,}74 \\ e_{12} &= 1 - 1{,}01 \times 1 - 0{,}72 = -0{,}73 \\ e_{13} &= 1 - 1{,}01 \times 1 - 0{,}72 = -0{,}73 \end{aligned}$$

Součet těchto odchylek je $\Sigma e_i = 0{,}04$ a součet čtverců $\Sigma e_i^2 = 6{,}2992$. Průměrná čtvercová odchylka residuí pak

bude $s_{xy} = \sqrt{\frac{\Sigma e_i^2}{r}} = \sqrt{0{,}4845} = 0{,}22$. V jakém poměru jsou k ní jednotlivá residua, je patrno z grafického znázornění v obr. 11.



Obr. 11. Meze dvojnásobné průměrné čtvercové odchylky residuí.

Najdeme nyní obecným řešením normálních rovnic konstanty přímky odhadu. Máme-li r dvojic pozorování, dostáváme r rovnic

$$\begin{aligned} y_1 &= ax_1 + b \\ y_2 &= ax_2 + b \\ &\cdots\cdots\cdots \\ y_r &= ax_r + b \end{aligned}$$

kde a, b jsou neznámé konstanty a x_i, y_i dostáváme z měření hodnot znaků. Normální rovnice pak jsou

$$\begin{aligned} \Sigma y &= a\Sigma x + rb, \\ \Sigma xy &= a\Sigma x^2 + b\Sigma x, \end{aligned}$$

jejichž řešením dostaneme pro konstanty

$$a = \frac{r\Sigma xy - \Sigma x\Sigma y}{r\Sigma x^2 - (\Sigma x)^2}, \qquad b = \frac{\Sigma y\Sigma x^2 - \Sigma x\Sigma xy}{r\Sigma x^2 - (\Sigma x)^2}, \quad (64)$$

což jsou výrazy složené z hlavních součtů hodnot znaků, jejich čtverců a podvojných součinů.

Podobně najdeme obecný výraz pro čtverec průměrné čtvercové odchylky residuí, neboť máme obecně r rovnic tvaru

$$\begin{aligned} e_i^2 &= [y_i - (ax_i + b)]^2 = \\ &= y_i^2 + a^2x_i^2 + b^2 + 2abx_i - 2ax_iy_i - 2by_i, \end{aligned}$$

které sečteme pro všechna i a dostaneme

$$\begin{aligned} \Sigma e^2 = \Sigma y^2 &+ (a^2\Sigma x^2 + 2ab\Sigma x + rb^2) - \\ &- 2a\Sigma xy - 2b\Sigma y. \end{aligned} \quad (65)$$

Násobíme-li první normální rovnici konstantou b, druhou a

a sečteme je, bude

$$a^2\Sigma x^2 + 2ab\Sigma x + rb^2 = a\Sigma xy + b\Sigma y.$$

Můžeme tedy dosaditi do rovnice (65) za výraz v kulaté závorce, takže

$$\Sigma e^2 = \Sigma y^2 + a\Sigma xy + b\Sigma y - 2a\Sigma xy - 2b\Sigma y = \\ = \Sigma y^2 - a\Sigma xy - b\Sigma y,$$

a tedy

$$s_{xy}{}^2 = \frac{\Sigma e^2}{r} = \frac{\Sigma y^2 - a\Sigma xy - b\Sigma y}{r}. \qquad (66)$$

Známe-li tudíž základní součty, můžeme bez námahy napsati rovnici přímky odhadu a průměrnou čtvercovou odchylku residuí. Použijeme-li výrazů (64), nemusíme vypisovati ani rovnice základní ani normální. Rovnici přímky odhadu vyjádříme ještě pomocí odchylek od průměrů. Podle rovnice I, (5) můžeme psáti

$$\Sigma\xi^2 = \Sigma x^2 - \frac{1}{r}(\Sigma x)^2,$$

při čemž opět vynecháváme index i, kde $\xi = x - \bar{x}$,

takže

$$\Sigma x^2 = \Sigma\xi^2 + \frac{1}{r}(\Sigma x)^2 \qquad (67)$$

a obdobně pro proměnnou y platí

$$\Sigma y^2 = \Sigma\eta^2 + \frac{1}{r}(\Sigma y)^2, \qquad (68)$$

kde $$\eta = y - \bar{y}.$$

Odvodíme snadno podobný výraz pro součet součinů $\Sigma\xi\eta$, neboť

$$\Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - \bar{x}\Sigma y - \bar{y}\Sigma x + r\bar{x}\bar{y} = \\ = \Sigma xy - \frac{1}{r}\Sigma x\Sigma y - \frac{1}{r}\Sigma y\Sigma x + \frac{1}{r}\Sigma x\Sigma y,$$

takže

$$\Sigma\xi\eta = \Sigma xy - \frac{1}{r}\,\Sigma x\Sigma y,$$

a tedy

$$\Sigma xy = \Sigma\xi\eta + \frac{1}{r}\,\Sigma x\Sigma y. \tag{69}$$

Dosadíme-li nyní výrazy (67), (69) do (64), dostaneme

$$a = \frac{r(\Sigma\xi\eta + \frac{1}{r}\Sigma x\Sigma y) - \Sigma x\Sigma y}{r(\Sigma\xi^2 + \frac{1}{r}\,(\Sigma x)^2) - (\Sigma x)^2} = \frac{\Sigma\xi\eta}{\Sigma\xi^2}. \tag{70}$$

Konstantu b můžeme psáti podle první normální rovnice

$$b = \bar{y} - a\,\bar{x},$$

takže rovnice přímky odhadu je

$$y = \frac{\Sigma\xi\eta}{\Sigma\xi^2}\,x + \bar{y} - \bar{x}\,.\,\frac{\Sigma\xi\eta}{\Sigma\xi^2}. \tag{71}$$

Obdobně vyjádříme průměrnou čtvercovou odchylku residuí dosazením (68) a (69) do (66), takže

$$s_{xy}{}^2 = \frac{1}{r}\left[\Sigma\eta^2 + \frac{1}{r}\,(\Sigma y)^2 - a\,(\Sigma\xi\eta + \frac{1}{r}\,\Sigma x\Sigma y) - \right.$$
$$\left. - (\bar{y} - a\,\bar{x})\,\Sigma y\right].$$

Vzhledem k tomu, že

$$\bar{x} = \frac{\Sigma x}{r} \quad \text{a} \quad \bar{y} = \frac{\Sigma y}{r},$$

zruší se čtyři členy a zůstane

$$s_{xy}{}^2 = \frac{1}{r}\,(\Sigma\eta^2 - a\Sigma\xi\eta),$$

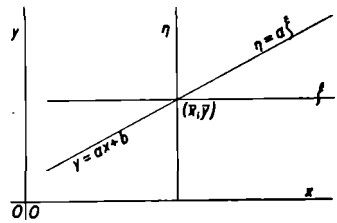
takže budeme psáti

$$s_{xy} = \sqrt{\frac{\Sigma\eta^2}{r} - \frac{(\Sigma\xi\eta)^2}{r\Sigma\xi^2}}. \qquad (72)$$

Velmi se zjednoduší tvar rovnice přímky odhadu, posuneme-li počátek souřadnic do bodu $(\bar{x}, \bar{y})$, který leží na této přímce, jak se snadno přesvědčíme, dosadíme-li tyto hodnoty do (71) za x resp. y. — Provedeme-li v (71) substituci $x = \bar{x} + \xi$, $y = \bar{y} + \eta$, dostaneme transformovanou rovnici přímky odhadu

$$\eta = \frac{\Sigma\xi\eta}{\Sigma\xi^2}\,\xi = a\xi. \qquad (73)$$

Výsledek této transformace, která spočívá jen v posunutí pravoúhlé soustavy souřadnic, je znázorněn v obr. 12.



Obr. 12. Rovnice přímky odhadu v původních proměnných a v odchylkách od průměrů.

Vyjádřili jsme konstantu a, jakož i průměrnou čtvercovou odchylku s_{xy}^2 pomocí součinů a čtverců odchylek hodnot znaků od průměrů. Výpočet skutečný lze zjednodušiti podobně jako při jedné proměnné užitím vhodně zvoleného počátku (I, str. 35). Budeme pak míti místo proměnné x novou proměnnou $v = x - x_0$. Jsou-li hodnoty x veliké, dosáhneme odečtením zatímního průměru, jak jsme číslo x_0 také nazvali, čísel menších. Víme pak, že odchylky hodnot v od jejich průměru jsou tytéž jako příslušné odchylky x od jejich průměru, neboť $\bar{v} = \bar{x} - x_0$, a tedy $v - \bar{v} = x - \bar{x}$. Totéž platí pro druhou proměnnou $w = y - y_0$.

Vzhledem k tomu můžeme tedy psáti výraz (70) pro konstantu a

$$a = \frac{\Sigma\xi\eta}{\Sigma\xi^2} = \frac{\Sigma(v-\bar{v})\,(w-\bar{w})}{\Sigma(v-\bar{v})^2} = \frac{\Sigma vw - \frac{1}{r}\,\Sigma v\Sigma w}{\Sigma v^2 - \frac{1}{r}\,(\Sigma v)^2}, \quad (74)$$

$$b = \bar{y} - a\bar{x} = \bar{w} + y_0 - a\,(\bar{v} + x_0).$$

Průměrnou čtvercovou odchylku pak můžeme psáti podle (72)

$$s_{xy}{}^2 = \frac{\Sigma\eta^2}{r} - \frac{(\Sigma\xi\eta)^2}{r\Sigma\xi^2} = \frac{\Sigma(w-\bar{w})^2}{r} - \frac{[\Sigma(v-\bar{v})(w-\bar{w})]^2}{r\Sigma(v-v)^2} =$$

$$= \frac{\Sigma w^2 - \frac{1}{r}\,(\Sigma w)^2}{r} - \frac{[\Sigma vw - \frac{1}{r}\,\Sigma v\Sigma w]^2}{r\,[\Sigma v^2 - \frac{1}{r}\,(\Sigma v)^2]}.$$

Tím jsme dostali výrazy, v nichž se vyskytují jen součty odchylek hodnot proměnných od zatímních průměrů.

(5,3,1) Příklad. Předpokládejme, že mezi třemi proměnnými x, y, z, platí lineární vztah $y = a_0 + a_1x + a_2z$. Hodnoty y jsou odhadovány na základě hodnot proměnných x a z. Jest najíti příslušnou rovnici roviny odhadu.

Máme celkem r rovnic z pozorování a potřebujeme k určení tří konstant tři normální rovnice. Můžeme nejprve ukázati, že stačí dvě normální rovnice, vyjádříme-li rovnici roviny odhadu pomocí odchylek od průměrů. Potom platí vztah

$$\eta = a'_0 + a_1\xi + a_2\zeta$$

a sestrojíme-li normální rovnice, dostaneme první sečtením všech r rovnic odvozených podle pozorování

$$\Sigma\eta = ra'_0 + a_1\Sigma\xi + a_2\Sigma\zeta; \quad (75)$$

další rovnice dostaneme, vynásobíme-li každou rovnici ξ resp. ζ a sečteme, takže bude

$$\Sigma\xi\eta = a'_0\Sigma\xi + a_1\Sigma\xi^2 + a_2\Sigma\xi\zeta,$$
$$\Sigma\zeta\eta = a'_0\Sigma\zeta + a_1\Sigma\xi\zeta + a_2\Sigma\zeta^2.$$

Poněvadž součty odchylek od průměru jsou rovny nule, bude $\Sigma\xi = \Sigma\eta = \Sigma\zeta = 0$, takže první rovnice se redukuje na $r \,.\, a'_0 = 0$, čili první konstanta $a'_0 = 0$ a zbývají dvě normální rovnice

$$\Sigma\xi\eta = a_1\Sigma\xi^2 + a_2\Sigma\xi\zeta,$$
$$\Sigma\zeta\eta = a_1\Sigma\xi\zeta + a_2\Sigma\zeta^2,$$

z nichž snadno vypočítáme a_1 a a_2 a rovnice roviny odhadu bude

$$\eta = a_1\xi + a_2\zeta.$$

(6,1) Koeficient korelace. Výraz pro průměrnou čtvercovou odchylku (72) upravíme dále tím, že vytkneme $\sigma_y^2 = \frac{\Sigma\eta^2}{r}$; tak dostaneme

$$s_{xy}^2 = \frac{\Sigma\eta^2}{r} - \frac{(\Sigma\xi\eta)^2}{r\Sigma\xi^2} = \sigma_y^2 \left\{1 - \frac{(\Sigma\xi\eta)^2}{\Sigma\xi^2\Sigma\eta^2}\right\},$$

čili

$$s_{xy}^2 = \sigma_y^2 (1 - r_{xy}^2), \tag{76}$$

kde klademe

$$r_{xy} = \frac{\Sigma\xi\eta}{\sqrt{\Sigma\xi^2\,\Sigma\eta^2}} \tag{77}$$

a tento výraz se nazývá koeficient korelace mezi x a y. Odvodil jej Bravais a jeho teorii korelace propracoval pak zvláště K. Pearson.

Je patrno, že můžeme koeficient korelace psáti také v tvaru

$$r_{xy} = \frac{\Sigma\xi\eta}{r\sigma_x\sigma_y}. \tag{78}$$

Průměr součinů dvou proměnných měřených od jejich průměrů

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{r} = \frac{\Sigma\xi\eta}{r}$$

se nazývá také jejich kovariance.

Výraz pro koeficient korelace mezi x a y je týž jako pro koeficient korelace mezi y a x, neboť je vzhledem k oběma proměnným zcela symetrický. Nezáleží tedy na tom, která by byla považována za odvislou a která za neodvislou proměnnou. Proto také nezáleží v symbolu koeficientu korelace na pořadí indexů, čili $r_{xy} = r_{yx}$.

Z rovnice (76) vidíme, že $s_{xy} = 0$, je-li $r_{xy} = \pm 1$, což znamená, že všechny residuální odchylky jsou rovny nule, čili všechny hodnoty y padnou na přímku odhadu a je to tedy případ úplné korelace mezi znaky x a y. Případ $s_{xy} = \sigma_y$ nastává, když $r_{xy} = 0$; jeho významu porozumíme, napíšeme-li rovnici přímky odhadu v novém tvaru. Podle (71) jest

$$y = \frac{\Sigma\xi\eta}{\Sigma\xi^2}\, x + \bar{y} - \bar{x}\frac{\Sigma\xi\eta}{\Sigma\xi^2},$$

kde můžeme psáti podle (78) $\Sigma\xi\eta = r\sigma_x\sigma_y r_{xy}$. Rovnice přímky odhadu pak bude

$$y = r_{xy}\frac{\sigma_y}{\sigma_x}\, x + \bar{y} - \bar{x}\frac{\sigma_y}{\sigma_x}\, r_{xy} \tag{79}$$

a pro $r_{xy} = 0$ se redukuje na $y = \bar{y}$, což znamená, že pro jakoukoliv hodnotu x bude vždy nejlepší hodnotou y průměr $\bar{y}$. Přímkou odhadu je zde rovnoběžka s osou x. Mezi znaky x a y není vázanosti. Abychom tedy změřili těsnost lineárního vztahu mezi dvěma znaky, počítáme koeficient korelace.

(6,2) Různé tvary koeficientu korelace. Není-li rozsah pozorovaného souboru větší než 50 a pozorované hodnoty proměnných x a y nejsou příliš velké, je výhodno počítati koeficient korelace podle výrazu

$$r_{xy} = \frac{\Sigma\xi\eta}{\sqrt{\Sigma\xi^2\,\Sigma\eta^2}} = \frac{r\Sigma xy - \Sigma x\Sigma y}{\sqrt{[r\Sigma x^2 - (\Sigma x)^2]\,[r\Sigma y^2 - (\Sigma y)^2]}}. \tag{80}$$

Jsou-li však hodnoty x, y velké, zjednoduší se výpočet metodou vhodně zvoleného počátku, takže se od každé pro-

měnné odečítá zatímní průměr. Pak se obdobně jako v (74) odvodí

$$r_{xy} = \frac{\Sigma(v-\bar{v})(w-\bar{w})}{\sqrt{\Sigma(v-\bar{v})^2\,\Sigma(w-\bar{w})^2}} = \tag{81}$$

$$= \frac{r\Sigma vw - \Sigma v\Sigma w}{\sqrt{[r\Sigma v^2-(\Sigma v)^2]\,[r\Sigma w^2-(\Sigma w)^2]}}.$$

Další tvar dostáváme z rovnice (78).

$$r_{xy} = \frac{\Sigma\xi\eta}{r\sigma_x\sigma_y} = \frac{\Sigma(v-\bar{v})(w-\bar{w})}{r\sigma_x\sigma_y} = \frac{\Sigma vw - r\bar{v}\bar{w}}{r\sigma_v\sigma_w}, \tag{82}$$

neboť $\sigma_x = \sigma_v$ a $\sigma_y = \sigma_w$;

řešíme-li rovnici (76) podle r_{xy}^2, dostaneme

$$r_{xy}^2 = 1 - \frac{s_{xy}^2}{\sigma_y^2} = 1 - \frac{\Sigma\eta^2 - a\Sigma\xi\eta}{\Sigma\eta^2}. \tag{83}$$

Dále můžeme vyjádřiti koeficient korelace pomocí rozdílů odchylek hodnot znaků od jejich průměrů, tedy $\xi - \eta$, stanovíme-li jejich průměrnou čtvercovou odchylku

$$\sigma_d^2 = \frac{\Sigma(\xi-\eta)^2}{r} = \frac{\Sigma\xi^2}{r} - \frac{2\Sigma\xi\eta}{r} + \frac{\Sigma\eta^2}{r} = \sigma_x^2 - \frac{2\Sigma\xi\eta}{r} + \sigma_y^2,$$

takže

$$\frac{2\Sigma\xi\eta}{r} = \sigma_x^2 + \sigma_y^2 - \sigma_d^2,$$

a odtud vzhledem ku (78)

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_d^2}{2\sigma_x\sigma_y}. \tag{84}$$

(6,3) Korelace pořadových čísel. V některých případech malých souborů se osvědčuje výpočet koeficientu korelace metodou pořadí. Nepoužívá se pak přímo hodnot proměnných, nýbrž se seřadí podle velikosti hodnoty jednoho znaku a hodnoty druhého znaku tak, že každá hodnota dostane

určité pořadí čili rang. Původní hodnoty proměnné x a y jsou tak nahrazeny dvěma řadami příslušných čísel pořadových i_x a i_y a jejich koeficient korelace se počítá. Jsou pak případy, kde není možno předpokládati, že mezi zkoumanými řadami pozorovaných čísel je s dostatečnou přibližností lineární vztah, ale čísla pořadí jejich se mu blíží; pak je koeficient korelace čísel pořadí spíše na místě a znamená také značnou úsporu práce.

Přidělení pořadových čísel by bylo zcela snadné, kdyby se každá hodnota proměnné vyskytovala jen jednou. Častěji máme však co činiti s případem, kde se jednotlivé hodnoty znaku vyskytují několikrát. Máme na př. podle velikosti seřazeným hodnotám znaku přiřadit pořadí

4,5	4,4	4,0	4,0	3,7	3,4	3,4	3,4	3,1	2,9	
1	2	3	4	5	6	7	8	9	10	(85)

Přiřadili bychom tedy hodnotě 4,0 pořadí 3 a 4, nebo hodnotě 3,4, která se vyskytuje třikrát, pořadí 6, 7, 8. Ale stejným původním číslům má odpovídat stejné číslo pořadí; proto jim obyčejně přiřazujeme průměr pořadových čísel, jež by jim patřila, kdyby byla vesměs od sebe různá. V tomto případě tudíž budou

1	2	3,5	3,5	5	7	7	7	9	10.

Výsledný tvar koeficientu korelace odvodíme z výrazu (80), uvážíme-li, že hodnotami proměnných je prvních r čísel celých podle (85), i když pak některá jsou nahrazena určitými průměry.

Pak tedy bude součet hodnot i_x roven součtu hodnot i_y a tedy roven součtu prvních r celých čísel.

$$\Sigma i_x = \Sigma i_y = \frac{r}{2}(r+1). \tag{86}$$

Rovněž je známo, že součet čtverců je

$$\Sigma i_x^2 = \Sigma i_y^2 = 1^2 + 2^2 + 3^2 + \ldots + r^2 = \frac{r(r+1)(2r+1)}{6}, \tag{87}$$

takže musíme ještě stanovit Σxy. Použijeme k tomu rozdílů pořadových čísel a dostáváme

$$d_1 = i_{x,1} - i_{y,1}, \text{ tudíž } d_1^2 = i_{x,1}^2 - 2i_{x,1}i_{y,1} + i_{y,1}^2,$$

odkud

$$2i_{x,1}i_{y,1} = i_{x,1}^2 + i_{y,1}^2 - d_1^2,$$

a součet všech členů bude

$$2\Sigma i_x i_y = \Sigma i_x^2 + \Sigma i_y^2 - \Sigma d^2 = 2\Sigma i_x^2 - \Sigma d^2,$$

$$\Sigma i_x i_y = \Sigma i_x^2 - \tfrac{1}{2}\Sigma d^2. \tag{88}$$

Stačí nyní dosaditi do výrazu (80) a pro koeficient korelace ϱ pořadových čísel plyne

$$\varrho = \frac{rr(r+1)(2r+1):6 - r\Sigma d^2:2 - r^2(r+1)^2:4}{\sqrt{[rr(r+1)(2r+1):6 - r^2(r+1)^2:4]^2}}$$

čili

$$\varrho = 1 - \frac{\Sigma d^2}{r(r+1)[(2r+1):3 - (r+1):2]} =$$

$$= 1 - \frac{6\Sigma d^2}{r(r+1)(4r+2-3r-3)},$$

takže je konečně

$$\varrho = 1 - \frac{6\Sigma d^2}{r(r^2-1)}. \tag{89}$$

Je patrno, že koeficient korelace pořadových čísel podle této t. zv. formule Spearmanovy lze v mnohých případech snadněji vypočítati, než podle formule Bravaisovy, ježto Σd^2 se snadno stanoví. Rozdíly pořadových čísel jsou obyčejně malá čísla, takže jejich čtverce se rychle určí a sčítají (viz tab. 15). Hodnoty koeficientů r_{xy} a ϱ jsou sobě obyčejně velmi blízké, ačkoliv tomu nemusí tak vždy býti, což si lze ukázati třeba jednoduchým příkladem dvou řad

x:	60,	50,	40,	30,	10
y:	100,	98,	97,	3,	1.

Koeficient korelace pořadových čísel dává těsnost vztahu úplnou $\varrho = 1$, kdežto r_{xy} se jedné nerovná.

Mezi oběma koeficienty korelace platí vztah odvozený za jistých předpokladů (t. zv. normální korelace)

$$r_{xy} = 2 \sin\left(\frac{180°}{6} \cdot \varrho\right). \tag{90}$$

Přímku odhadu v případě korelace pořadových čísel dostaneme, dosadíme-li do rovnic (64) výrazy (86), (87), (88), takže pak se zřetelem ku (89) je

$$a = \varrho, \quad b = \tfrac{1}{2}(1 - \varrho)(r + 1)$$

a rovnice přímky odhadu tedy bude

$$y = \varrho x + \tfrac{1}{2}(1 - \varrho)(r + 1).$$

Jednoduchý odhad korelace mezi dvěma znaky pomocí pořadových čísel lze provésti podle formule navržené rovněž Spearmanem

$$\varrho_0 = 1 - \frac{s}{m},$$

kde s značí součet kladných rozdílů mezi pořadovými čísly a $m = \dfrac{r^2 - 1}{6}$. Za předpokladu normálního rozdělení četností platí pak vztah

$$r_{xy} = 2 \cos \frac{180°}{3}(1 - \varrho_0) - 1.$$

(6,4) Schema pro výpočet koeficientu korelace z korelační tabulky. Také při výpočtu charakteristik dvojrozměrného rozdělení četností se doporučuje zachovávati určitý stálý postup, zvláště při procvičování látky. Rozvrhneme tedy vhodně do formuláře rubriky, kterých je třeba k výpočtu koeficientu korelace podle rovnice (82) pro soubor nevelkého rozsahu a malého počtu třídních intervalů. Pro hodnoty proměnné x, která může znamenati na př. měsíční cenový index, pozorovaný po tři léta, uvedeme třídní znaky

a rovněž pro hodnoty proměnné y, která může znamenati třeba výrobu surového železa v milionech tun. Tříděním podle těchto dvou znaků dostaneme korelační tabulku 14. Délka třídních intervalů proměnné x je $h_1 = 0{,}5$, kdežto

Tabulka 14.

y \ x	7,7	8,2	8,7	9,2	9,7	n_y	w	wn_k	w^2n_k	s_k	s_kw	s_k^2	$\frac{s_k^2}{n_k}$
10	7	3				10	—2	—20	40	—17	34	289	28,90
14	10	8	2	2	1	23	—1	—23	23	—24	24	576	25,04
18	8	21	14	3	1	47	0	—	—	—32	—	1024	21,79
22		9	26	7	5	47	1	47	47	8	8	64	1,36
26			4	16	1	21	2	42	84	18	36	324	15,43
30				4	4	8	3	24	72	12	36	144	18,00
n_x	25	41	46	32	12	156		70	266		138		110,52
v	—2	—1	0	1	2								
vn_i	—50	—41	—	32	24	—35							
v^2n_i	100	41	—	32	48	221							
s_i	—24	— 5	32	49	18								
s_iv	48	5	—	49	36	138							
s_i^2	576	25	1024	2401	324								
$\frac{s_i^2}{n_i}$	23,04	0,61	22,26	75,03	27,00	147,94							

$$\bar{w} = \frac{70}{156} = 0{,}4487h_2,$$

$$\sigma_w^2 = \frac{266}{156} - \bar{w}^2 = 1{,}5038,$$

$$\sigma_w = 1{,}2263h_2,$$

$$\bar{v} = \frac{-35}{156} = -0{,}2244h_1,$$

$$\sigma_v^2 = \frac{221}{156} - \bar{v}^2 = 1{,}3663,$$

$$\sigma_v = 1{,}1689h_1,$$

$$\frac{\Sigma vw}{r} = \frac{138}{156} = 0{,}8846h_1h_2,$$

$$r_{xy} = \frac{1}{\sigma_v\sigma_w}\left(\frac{\Sigma vw}{r} - \bar{v}\bar{w}\right) = 0{,}687.$$

délka třídních intervalů proměnné y je $h_2 = 4{,}0$. Jinak vyžaduje výkladu jen sloupec $s_{\frac{1}{2}}$. Každý člen v něm je algebraickým součtem odchylek v násobených příslušnou četností jeho řádku. Tak dostaneme první číslo -17 jako součet $(-2) \cdot 7 + (-1) \cdot 3 = -17$, neboť -2 resp. -1 jsou odchylky těch hodnot znaku, jimž přísluší v tomto prvním řádku korelační tabulky četnosti 7, resp. 3. Číslo 18 v témž sloupci dostaneme jako součet $0 \cdot 4 + 1 \cdot 16 + 2 \cdot 1 = 18$. Zcela obdobně vzniká sloupec $s_{\frac{1}{1}}$. V případě zpracování tabulky rozdělení četností o větším počtu tříd je vhodno zapisovati jednotlivé součiny do rohů pole příslušné četnosti.

(6,5) Výpočet koeficientu korelace z řad hodnot dvou znaků. Provedeme si nyní výpočet koeficientu korelace v případě, že počet prvků souboru je malý, takže dvojice hodnot znaků stanovené na těchto prvcích nebyly sestaveny do tabulky o dvojím vstupu. Máme tedy 15 prvků, na nichž byly stanoveny hodnoty znaků x a y zapsané v tab. 15.

Tabulka 15.

i	x	y	i_x	i_y	d_i	d_i^2
1	19	25	10	10,5	—0,5	0,25
2	73	100	1	1,5	—0,5	0,25
3	31	50	6,5	5	+1,5	2,25
4	8	10	14,5	14,5	0,0	0,00
5	54	50	3	5	—2,0	4,00
6	71	100	2	1,5	+0,5	0,25
7	22	25	9	10,5	—1,5	2,25
8	8	25	14,5	10,5	+4,0	16,00
9	33	50	5	5	0,0	0,00
10	31	50	6,5	5	+1,5	2,25
11	41	50	4	5	—1,0	1,00
12	23	25	8	10,5	—2,5	6,25
13	10	25	12	10,5	+1,5	2,25
14	10	25	12	10,5	+1,5	2,25
15	10	10	12	14,5	—2,5	6,25
	444	620	120,0	120,0		45,50

Tabulka 16.

x^2	y^2	xy	$(x+y)^2$
361	625	475	1936
5329	10000	7300	29929
961	2500	1550	6561
64	100	80	324
2916	2500	2700	10816
5041	10000	7100	29241
484	625	550	2209
64	625	200	1089
1089	2500	1650	6889
961	2500	1550	6561
1681	2500	2050	8281
529	625	575	2304
100	625	250	1225
100	625	250	1225
100	100	100	400
19780	36450	26380	108990

Vypočítáme napřed koeficient korelace pořadových čísel, která jsou v tabulce zapsána v čtvrtém a pátém sloupci. Najdeme nejprve největší hodnotu znaku x, které přiřadíme pořadové číslo 1, nejbližší nižší hodnotě číslo 2 atd. Znak y má dvě stejné největší hodnoty 100, takže každému z nich přiřadíme průměrné číslo $1{,}5 = \frac{1}{2}(1 + 2)$. Po této hodnotě je nejbližší 50, která se vyskytne pětkrát; bude tedy míti pořadové číslo $5 = \frac{1}{5}(3 + 4 + 5 + 6 + 7)$ a podobně se postupuje dále. Utvoříme pak rozdíly, příslušných pořadových čísel a jejich čtverce. Podle rovnice (89) potom dostáváme

$$\varrho = 1 - \frac{6 \,.\, 45{,}50}{15\,(15^2 - 1)} = 1 - 0{,}081 = + 0{,}919.$$

Pro srovnání vypočítáme také koeficient korelace r_{xy} podle rovnice (80). Příslušná čísla jsou v tabulce 16; pro kontrolu je připojen poslední sloupec, abychom zjistili, že

$$(108990 - 19780 - 36450) : 2 = 26\,380,$$

$$r_{xy} = \frac{15 \,.\, 26380 - 444 \,.\, 620}{\sqrt{(15 \,.\, 19780 - 444^2)\;\;(15 \,.\, 36450 - 620^2)}},$$

$$r_{xy} = + 0{,}947.$$

Podle vztahu (90) bychom dostali

$$r_{xy} = 2 \sin 27^\circ 34{,}2' = 0{,}926.$$

Odchylkami, které jsme dostali mezi hodnotami ϱ a r_{xy} se budeme zabývati později.

Když vykládáme smysl dosažených výsledků v určitých případech, mějme na paměti, že statistické vztahy představují statistické pravidelnosti; jejich výpovědi platí jen pro zkoumaný statistický soubor jako celek, nikoliv pro jednotlivé prvky souboru.

(6,5,I) Příklad 1. Jest vypočítati koeficienty korelace pro každý ze tří souborů tab. 10.

Soubor č. 1. K oběma řadám čísel x a y si sestavíme ještě další tři sloupce.

v \ w	−4	−3	−2	−1	0	1
−6			1 12		0	1 6
−5	1 20			3 5	0	
−4	1 16	1 12	2 8	2 4	0	1 4
−3			2 6	3 3	0	5 3
−2		1 6	1 4	⑤ 2	0	[6] 2
−1				7 1	0	5 1
0	0	0	0	0	0	0
1				2 1	0	7 1
2				[1] 2	0	⑨ 2
3					0	5 3
4					0	3 4
5					0	
6					0	
7					0	
wm	−8	−6	−12	−25	−51	47
w^2n	32	18	24	25		47

17.

2	3	4	5	6	vn	v^2n
					—12	72
					—20	100
					—32	128
1 6					—54	162
1 4					—46	92
[2] 2	1 3				—31	31
0	0	0	0	0	—195	
⑦ 2	1 3	2 4		1 6	26	26
5 4	1 6				30	60
9 6	2 9	1 12			60	180
3 8	1 12		1 20		32	128
2 10		1 20		1 30	20	100
		1 24	1 30		18	108
1 14					7	49
74	30	20	10	12	193	1236
148	90	80	50	72	586	

x^2	y^2	xy
64	81	72
16	64	32
49	25	35
49	81	63
1	36	6
4	16	8
36	1	6
25	9	15
9	1	3
16	36	24
81	16	36
9	36	18
1	64	8
360	466	326

Dostáváme tedy součty

$$\Sigma x = 60,\ \Sigma y = 70,\ \Sigma x^2 = 360,$$
$$\Sigma y^2 = 466,\ \Sigma xy = 326,$$

z nichž vyplývají hodnoty charakteristik

$$\bar{x} = 4{,}62,\ \bar{y} = 5{,}38,$$
$$\sigma_x^2 = \tfrac{360}{13} - 4{,}62^2 = 6{,}3479,$$
$$\sigma_y^2 = \tfrac{466}{13} - 5{,}38^2 = 6{,}9018,$$
$$\sigma_x = 2{,}52,\ \sigma_y = 2{,}63.$$

Koeficient korelace bude podle rovnice (80)

$$r_{xy} = +\,0{,}03.$$

Soubor č. 2. Příslušné součty jsou

$$\Sigma x = 60,\ \Sigma y = 70,\ \Sigma x^2 = 360,$$
$$\Sigma y^2 = 468,\ \Sigma xy = 407,$$

takže

$$\bar{x} = 4{,}62, \qquad \bar{y} = 5{,}38,$$
$$\sigma_x^2 = 6{,}3479, \qquad \sigma_x = 2{,}52,$$
$$\sigma_y^2 = 7{,}0556, \qquad \sigma_y = 2{,}64.$$

Koeficient korelace podle rovnice (80) pak je $x_{xy} = +\,0{,}97$.

Soubor č. 3. Potřebné součty jsou tytéž jako pro soubor č. 2, až na $\Sigma xy = 238$, takže také charakteristiky jsou tytéž až na koeficient korelace, který tu je

$$r_{xy} = -\,0{,}98.$$

Příklad 2. Jest stanoviti koeficient korelace r_{xy} pro soubor daný v tabulce 11.

Můžeme použíti postupu podaného v tab. 14, nebo jej poněkud pozměníme. Zavedeme si stejně nové proměnné v, w, měřené v příslušné délce intervalu jako jednotce, a od vhodně zvoleného počátku. Tak zvolíme pro v počáteční hodnotu $v_0 = 5{,}4$ a pro druhou proměnnou $w_0 = 3{,}1$. Do příslušného pole tabulky zaznamenáme součin $v \,.\, w$ a dostaneme tab. 17.

Ve vyznačeném kříži jsou součiny $v \,.\, w$ rovny nule, protože je tam aspoň jeden součinitel roven nule a v poli, kde se oba pásy překrývají, jsou oba součinitelé rovny nule. Ostatní hodnoty součinů vepíšeme do pravého dolního rohu každého pole. Při tom vezmeme v úvahu dále, že křížem je celá tabulka rozdělena na čtyři oblasti tak, že v levé horní a pravé dolní jsou hodnoty součinů kladné, kdežto v ostatních dvou záporné.

Abychom nyní vypočítali $\Sigma v \,.\, w$, sestavíme si pomocnou tabulku 18, v níž do 1. sloupce sestavíme podle velikosti

Tabulka 18.

vw (1)	$n+$ (2)	$n-$ (3)	Alg. součet n (4)	$v \,.\, w \,.\, n$ (5)
1	14	7	7	7
2	18	9	9	18
3	9	6	3	9
4	13	2	11	44
5	9	0	3	15
6	14	2	12	72
8	5		5	40
9	2		2	18
10	2		2	20
12	4		4	48
14	1		1	14
16	1		1	16
20	3		3	60
24	1		1	24
30	2		2	60
	92	26		465

všechny hodnoty součinů $v \,.\, w$, které se vyskytují. Do 2. sloupce zapíšeme součty četností kladných oblastí postupně z těch polí, kde je $v \,.\, w = 1, 2, 3, \ldots$, do 3. sloupce obdobné součty z oblastí záporných. Do 4. sloupce zapisujeme algebraický součet sloupce 2. a 3.; v 5. sloupci pak máme vynásobený sloupec 4. sloupcem 1.

Tak dostáváme na př. $18 = 5 + 6 + 7$ (čísla označená v tab. 17 kroužky) a $-9 = -6 - 2 - 1$ (čísla označená čtverečky).

Z těchto tabulek dostáváme již všechna potřebná čísla k výpočtu charakteristik a koeficientu korelace.

$$\bar{v} = (193 - 195) : 200 = -0{,}01,$$
$$\bar{w} = (193 - 51) : 200 = +0{,}71,$$
$$\sigma_v^2 = \frac{1\,2\,3\,6}{2\,0\,0} - 0{,}00001 = 6{,}1799,$$
$$\sigma_v = 2{,}486,$$
$$\sigma_w^2 = \frac{5\,8\,6}{2\,0\,0} - 0{,}5041 = 2{,}4259,$$
$$\sigma_w = 1{,}557,$$
$$\frac{\Sigma v\,wn}{r} - \bar{v} \cdot \bar{w} = \frac{4\,6\,5}{2\,0\,0} - 0{,}01 \cdot 0{,}71 = 2{,}3179,$$

takže

$$r_{xy} = \frac{2{,}3179}{2{,}486 \times 1{,}557} = 0{,}60.$$

Příklad 3. Cena nějakého statku a poptávka po něm jsou ve vztahu. Budeme pozorovati tento vztah na určitém statku všeobecné spotřeby. Prvním znakem x bude cena statku v měnové jednotce, druhým znakem y bude počet kusů prodaných při dotyčné ceně na př. v milionech. Dostaneme dvě řady čísel, uvedené v tab. 19.

Tabulka 19.

x	y	x^2	y^2	xy
6	22	36	484	132
5,5	25	30,25	625	137,5
5	27	25	729	135
4,5	28,5	20,25	812,25	128,25
4	30	16	900	120
3	31	9	961	93
28,0	163,5	136,50	4511,25	745,75

$\bar{x} = 4{,}67$
$\bar{y} = 27{,}25$
$\sigma_x = 0{,}99$
$\sigma_y = 3{,}05$

Vidíme, že korelace je inversní, neboť nižším hodnotám znaku x odpovídají vyšší hodnoty znaku y. Zjistíme-li koeficient korelace, dostáváme

$$r_{xy} = -0{,}956.$$

Koeficient korelace nám sice dává obraz těsnosti vztahu, ale nevidíme z něho, zda určité změně hodnoty x odpovídá stejná změna hodnoty y nebo větší či menší. Tato okolnost vynikne, budeme-li týž vztah pozorovati na druhém statku, který není předmětem všeobecné spotřeby, nýbrž je statkem přepychovým. Dostaneme pozorované dvojice v tab. 20.

Tabulka 20.

x	y	x^2	y^2	xy
600	0,5	360000	0,25	300
550	0,8	302500	0,64	440
500	1,2	250000	1,44	600
450	2,0	202500	4,00	900
400	2,9	160000	8,41	1160
300	4,0	90000	16,00	1200
2800	11,4	1365000	30,74	4600

$\bar{x} = 466{,}7$
$\bar{y} = 1{,}90$
$\sigma_x = 98{,}4$
$\sigma_y = 1{,}23$

Koeficient korelace je

$$r_{xy} = -0{,}992.$$

Vidíme, že změna hodnoty znaku y v případě prvního statku není poměrně tak velká jako změna hodnoty znaku x, neboť pokles o 50% v hodnotě x vyvolává vzrůst přibližně o 40% proměnné y. V případě druhého statku však pokles hodnoty znaku x o polovinu způsobuje vzrůst na osminásobek v hodnotě znaku y. Poměr těchto změn studujeme pomůckami, jež jsou vyloženy v dalších odstavcích (7,1).

Úloha. Ačkoliv nemá logického smyslu počítati koeficient korelace mezi proměnnými, které jsou vázány jednoznačnou matematickou funkcí, jako na př. mezi x a $y = x^k$, přece je

zajímavo, že koeficient korelace mezi celými čísly 1, 2, 3, . . ., r a jejich čtverci 1^2, 2^2, 3^2, . . ., r^2 má při $r \to \infty$ hodnotu $r_{xy} =$ $= 0{,}968$, pro třetí mocniny 1^3, 2^3, . . ., r^3 při $r \to \infty$ je $r_{xy} =$ $= 0{,}9512$ tedy ještě menší. Ověřte si tyto výsledky a počítejte koeficienty korelace mezi posloupnostmi 1, 2, 3, . . ., r a 1^k, 2^k, 3^k, . . ., r^k pro několik hodnot $k = 1, 2, 3, \ldots$. Při tom použijte obecného vztahu

$$s_k = 1^k + 2^k + 3^k + \ldots + r^k = \frac{(r+1)^{k+1} - (r+1)}{k+1} -$$

$$- \frac{k}{2!} s_{k-1} - \frac{k(k-1)}{3!} s_{k-2} - \frac{k(k-1)(k-2)}{4!} s_{k-3} - \ldots - s_1.$$

Tudíž

$$s_0 = r, \quad s_1 = \frac{r(r+1)}{2}, \quad s_2 = \frac{r(2r+1)(r+1)}{6},$$

$$s_3 = \left[\frac{r(r+1)}{2}\right]^2, \quad s_4 = \tfrac{1}{30}\{r(r+1)(2r+1)(3r^2+3r-1)\} \text{ atd.}$$

(7,1) Koeficienty regrese. Budeme nyní považovati ze dvou proměnných jednu, třeba x za nezávislou a druhou y za závislou. Zjistíme-li koeficient korelace r_{xy}, ukazuje nám, jaká se jeví těsnost vztahu mezi proměnnými a nebývá snadné porozuměti jeho stupnici, jakož i pochopiti význam určité jeho hodnoty, na př. 0,65. Bývá často prospěšnější, můžeme-li dáti nějaký odhad pravděpodobné změny v y, pro nějakou danou změnu v x. Tak na př. pro tabulku 11 dostáváme, že změně v délce o 1 odpovídá průměrně změna v šířce o 0,39. Zjišťuje se tedy v první fázi korelační analysy těsnost vztahu a druhou fázi tvoří určení nejpravděpodobnějšího vztahu; k této fázi nyní přistoupíme.

Viděli jsme již, že k povaze vztahu mezi x a y se dostáváme výpočtem průměrů sloupců nebo řádků korelační tabulky. Se změnou proměnné x o jednotku nastává přibližně změna o 1,5 v druhé proměnné. Může to býti lineární vztah a pak jej můžeme vyjádřit rovnicí přímky odhadu $\eta = a\xi$ čili podle (79)

$$\eta = \frac{\sigma_y}{\sigma_x} r_{xy} \xi, \tag{91}$$

neboť

$$a = r_{xy}\frac{\sigma_y}{\sigma_x} = r_{xy}\sqrt{\frac{\Sigma\eta^2}{\Sigma\xi^2}}.$$

Povahu této přímky si můžeme objasniti na datech tabulky 14. Vztah mezi proměnnými je pro ně dán rovnicí

$$\eta = 0{,}687\left(\frac{4{,}9052}{0{,}5845}\right)\xi,$$

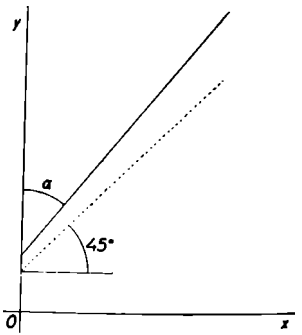
čili $\eta = 5{,}77\xi$. To znamená, že pro každou změnu o 1 v proměnné x nastává změna v y o 5,77.

Právě uvažovaná rovnice vyjadřuje vztah mezi x a y pomocí jejich odchylek od příslušných průměrů. Pro některé případy může býti vhodnějším napsati vztah pomocí původních hodnot pozorování. Pak bude příslušná rovnice

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x}), \quad (92)$$

čili

$$y - 19{,}79 = 5{,}77\ (x - 8{,}64),$$
$$y = 5{,}77x - 30{,}06.$$



Obr. 13. Galtonův graf.

Nakreslíme-li v souřadnicích proměnných x a y přímku regrese, pak její sklon udává poměr variací obou proměnných. Je-li tento poměr roven 1, bude míti regresní přímka sklon 45°. Je-li její sklon k horizontální ose větší než 45°, je tu větší poměrná změna v proměnné y než v x. Skutečný poměr variací je dán tangentou úhlu, který svírá regresní přímka s vertikální osou y, tedy tg α.

Uvedenému znázornění se také říká Galtonův graf.

Kdyby lineární vztah mezi oběma proměnnými byl perfektní, existovala by jedna přímka regresní. Je-li však vztah volný, takže $|r_{xy}| < 1$, dostáváme obdobnou úvahou jako

jsme odvodili rovnici (91) druhou přímku, kde považujeme y za nezávisle proměnnou a x za závislou proměnnou, jejíž rovnice bude

$$\xi = r_{xy} \frac{\sigma_x}{\sigma_y} \eta, \tag{93}$$

čili

$$x - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}). \tag{94}$$

Odvození regresních přímek jsme provedli zcela obecně jako přímek odhadu, ač se odvozují obyčejně přímo pro případ korelační tabulky. Odvození jejich pomocí koeficientu korelace je dovoleno jen když vztah mezi proměnnými je lineární. Rovnice (91) až (94) jsou rovnice přímek regresních, je-li regrese přesně lineární. Odchyluje-li se regrese od linearity buď v důsledku výběrových variací nebo skutečně svou povahou, dávají tyto rovnice nejlepší přímky regrese, které pozorovaná data připouštějí.

Můžeme se dívati na tyto rovnice buď a) jako na přímky odhadu individuálních hodnot y podle sdružených hodnot x a obráceně také odhadu hodnot x podle sdružených hodnot y; při tom jsou přímky stanoveny tak, že součet čtverců chyb odhadu je minimem nebo b) jako na přímky odhadu průměru hodnot y podle sdružených jednotlivých hodnot x a obráceně odhadu průměru hodnot x podle sdružených jednotlivých hodnot y; při tom opět součet čtverců chyb odhadu je minimem a stanoví se tak, že každý průměr se počítá tolikrát, kolik je prvků, z nichž byl stanoven. Je to tedy zase případ a), kde každá hodnota znaku y pro určité x byla zastoupena svým průměrem a pro druhou přímku každá hodnota x pro určité y byla zastoupena svým průměrem.

V rovnici regresní přímky znaku y vzhledem ku x (91) nebo (92) je koeficient, který znamená směrnici přímky

$$b_{21} = r_{xy} \frac{\sigma_y}{\sigma_x} \tag{95}$$

a nazývá se koeficient regrese y vzhledem ku x. Podobně v druhé přímce regresní (93) nebo (94)

$$b_{12} = r_{xy}\frac{\sigma_x}{\sigma_y} \tag{96}$$

je koeficient regrese x vzhledem ku y. Každá z těchto přímek prochází průměrem $(\bar{x}, \bar{y})$ celé korelační tabulky. Z rovnice (91) vidíme, že koeficient korelace je vyjádřen poměrem

$$\frac{\eta}{\sigma_y} : \frac{\xi}{\sigma_x} = r_{xy}.$$

Je také patrno, že koeficient korelace je geometrickým průměrem obou koeficientů regrese, neboť z rovnic (95) a (96) vyplývá

$$r_{xy} = \sqrt{b_{21} \cdot b_{12}}.$$

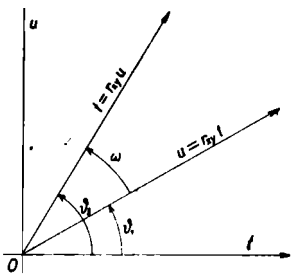
Vyjádříme-li pak rovnice regrese ve směrodatných proměnných $\frac{\xi}{\sigma_x} = t$, $\frac{\eta}{\sigma_y} = u$, dostáváme

$$u = r_{xy}t, \quad t = r_{xy}u, \tag{97}$$

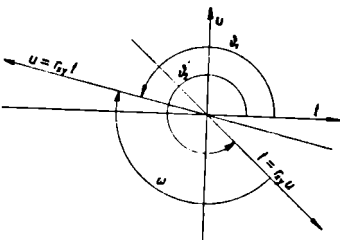
což jsou přímky odhadu vyjádřené v jednotkách směrodatných odchylek a pomocí koeficientu korelace. Jsou to abstraktní proměnné a nezávisí na jednotkách, v nichž byly měřeny původní proměnné x a y. Koeficient korelace je tedy tangentou úhlu, který svírá první přímka regrese s osou t a který označíme ϑ_1 a tangentou úhlu, který svírá druhá přímka s osou u, kdežto tangenta úhlu ϑ_2, který svírá druhá přímka regrese s osou t je $\frac{1}{r_{xy}}$. Úhel, který svírají obě přímky, označíme $\omega = \vartheta_2 - \vartheta_1$ a víme, že tedy

$$\operatorname{tg}\omega = \frac{\operatorname{tg}\vartheta_2 - \operatorname{tg}\vartheta_1}{1 + \operatorname{tg}\vartheta_1 \operatorname{tg}\vartheta_2} = \frac{\frac{1}{r_{xy}} - r_{xy}}{1 + 1} = \frac{1 - r_{xy}^2}{2r_{xy}}.$$

Poněvadž $1 - r_{xy}^2 \geqq 0$ závisí velikost úhlu ω na znaménku r_{xy}. Je-li $0 < r_{xy} < 1$ čili korelace kladná, bude $0 < \omega < R$ čili sevřený úhel bude ostrý. Je-li $-1 < r_{xy} < 0$ čili korelace záporná, je $R < \omega < 2R$. Pro $r_{xy} = \pm 1$, kdy je korelace perfektní, dostáváme $\operatorname{tg} \omega = 0$ čili $\omega = 0$. Obě regresní přímky splynou v jednu a to ve stejném smyslu je-li znaménko kladné čili při závislosti přímé, a v opačném smyslu, je-li znaménko záporné čili při závislosti nepřímé. Při $r_{xy} = 0$ svírají spolu obě regresní přímky úhel pravý a splývají tedy s osami t, u. V grafickém znázornění vyznačujeme šipkami



Obr. 14. Regresní přímky při $0 \leqq r_{xy} \leqq 1$.



Obr. 15. Regresní přímky při $0 \geqq r_{xy} \geqq -1$.

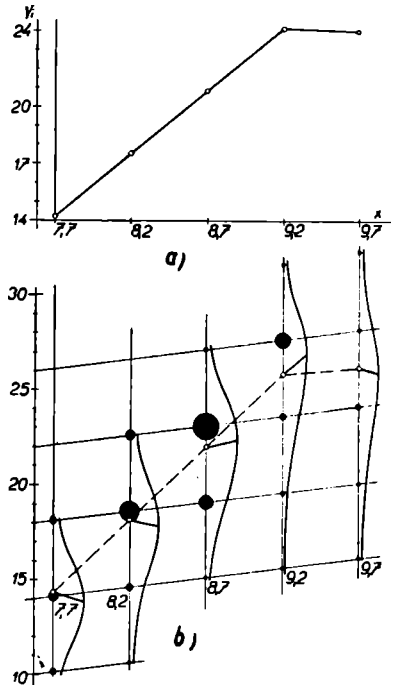
smysl přímek regresních, to jest směr rostoucích hodnot každého znaku. Případ současného růstu obou proměnných je znázorněn v obr. 14, kdežto případ nepřímé korelace je v obr. 15.

Znázorňujeme-li vztah mezi dvěma proměnnými, jejichž rozdělení četnosti je dáno korelační tabulkou, činíme tak pomocí průměrů. Tak na př. znázorníme regresní čáru tak, že hodnotám jedné proměnné x, kterou považujeme jako by za nezávislou, přiřadíme hodnoty průměrů sloupcových druhé proměnné; druhou regresní čáru dostaneme, považujeme-li proměnnou y za nezávislou a přiřazujeme jí řádkové průměry proměnné x.

Tak na př. pro rozdělení v tab. 14 máme

x_i	7,7	8,2	8,7	9,2	9,7
$\overline{y}_i$	14,2	17,5	20,8	24,1	24,0

Znázorníme-li jednotlivé body $(x_i, \overline{y}_i)$ a spojíme úsečkami, dostaneme čáry v obr. 16a, který podává jen informaci obsaženou v těchto bodech. Každý z těchto bodů je však průměrem několika hodnot, spadajících do tohoto sloupce, takže je jakousi oponou, v jejímž pozadí je určité rozdělení četností pozorovaného souboru. Představujeme si pak, že tomuto souboru odpovídá určitý základní soubor, který má své rozdělení četností v každém sloupci, jehož náhodným přiblížením je rozdělení pozorované. Znázorníme to tak, že v obr. 16b na každé pořadnici představující sloupec odpovídající určité hodnotě x, nakreslíme body odpovídající hodnotám y i s jejich vahami (velikostí teček) a připojíme rozdělení četností, znázorňující



Obr. 16. Regresní čára a její pozadí.

hypotetické rozdělení v příslušném sloupci základního souboru.

Vidíme z toho, že informace podávaná jednotlivými body představujícími průměry je podstatně doplněna rozptyly dotyčného sloupce.

(7,2) Koeficient determinace. Stanovili jsme průměrnou čtvercovou odchylku residuí, t. j. průměr čtverců odchylek měřených rovnoběžně s osou y od přímky odhadu čili od přímky regrese a uvedli jsme ji na tvar (76), z něhož řešením dostáváme pro r_{xy}^2

$$r_{xy}^2 = 1 - \frac{s_{xy}^2}{\sigma_y^2}.$$

Vidíme z toho, že čím je koeficient korelace větší, tedy bližší 1, tím je s_{xy} menší a naopak se blíží s_{xy} ku σ_y, když se hodnota r_{xy}^2 blíží k nule. Hodnota r_{xy}^2 se někdy nazývá koeficientem determinace, ježto měří procento variability hodnot odvislé proměnné určených z hodnot neodvislé proměnné. Můžeme ji psát také

$$r_{xy}^2 = \frac{\sigma_y^2 - s_{xy}^2}{\sigma_y^2},$$

odkud je zřejmo, že je to poměr rozdílu rozptylů $\sigma_y^2 - s_{xy}^2$ k rozptylu σ_y^2. Když s_{xy}^2 představuje rozptyl hodnot odvisle proměnné kolem přímky odhadu a σ_y^2 rozptyl těchto hodnot kolem celkového průměru, pak rozdíl $\sigma_y^2 - s_{xy}^2$ je výrazem té části rozptylu odvislé proměnné, která připadá na rozptyl způsobený neodvisle proměnnou.

(7,2,1) Příklad 1. Znázorněte graficky různou velikostí teček rozdělení četností uvedené v tabulce 14, průměry řádků a sloupců, jakož i obě přímky regresní.

Průměry řádků jsou

y_k	10	14	18	22	26	30
$\bar{x}_k$	7,85	8,18	8,36	8,78	9,13	9,45

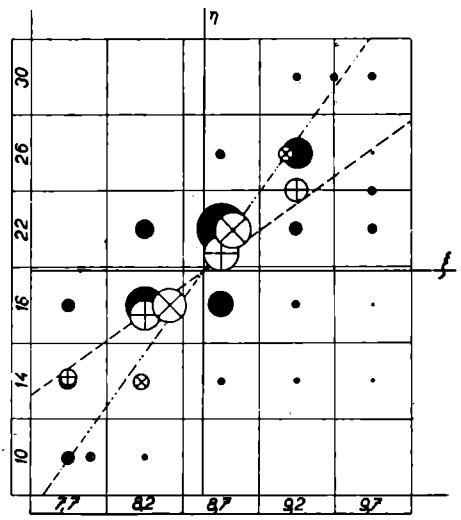
Průměry sloupců

x_i	7,7	8,2	8,7	9,2	9,7
$\bar{y}_i$	14,2	17,5	20,8	24,1	24,0

Rovnice regresních přímek jsou

$$\eta = 5{,}77\xi, \qquad \xi = 0{,}687\,\frac{0{,}5845}{4{,}9052}\,\eta.$$

Grafickým znázorněním polohy bodů, představujících jednotlivé páry hodnot si můžeme získat často potřebné objasnění, zda je splněna podmínka lineárního vztahu, nebo zda jen jednotlivé výbočující body ruší lineární průběh. Čím přesněji leží body u nějaké přímky, tím je r_{xy} blíže ± 1.



Obr. 17. Četnosti hodnot znaků a průměrů, jakož i přímky regrese.

Příklad 2. Odvoďte rovnici přímky regrese hodnot y vzhledem k x pro průměry, jejichž váhy se rovnají četnosti příslušného sloupce. Píšeme rovnici přímky, která se přimyká bodům, jejichž souřadnice jsou x_i, $\bar{y}_i$ tak, že součet čtverců odchylek rovnoběžných s osou y je nejmenší, ve tvaru $\bar{y}_i = ax_i + b$. Četnost sloupců je n_i, a přiřazena jako váha příslušného průměru značí, že každá pořadnice náležející v korelační tabulce určité hodnotě x_i je zastoupena průměrem všech pořadnic patřících k téže úsečce x_i. Potom má býti splněna podmínka, aby součet čtverců odchylek $\bar{y}_i - ax_i - b$ tedy

$$f(a, b) = \sum_i n_i (\bar{y}_i - ax_i - b)^2$$

byl minimem. K tomu musejí býti splněny především podmínky

$$\frac{\partial f}{\partial b} = \sum_i n_i (\bar{y}_i - ax_i - b) = 0,$$

$$\frac{\partial f}{\partial a} = \sum_i n_i (\bar{y}_i - ax_i - b)\, x_i = 0.$$

Jsou-li hodnoty proměnné x v korelační tabulce $x_1, x_2, \ldots, x_i, \ldots, x_l$ a hodnoty druhé proměnné $y_1, y_2, \ldots, y_k, \ldots, y_m$ pak četnost dvojice hodnot x_i, y_k označíme $n_{i,k}$. Korelační tabulka obsahuje l sloupců a m řádků. Součet četností k-tého řádku bude $\sum_{j=1}^{l} n_{i,k} = n_k$ a součet četností i-tého sloupce $\sum_{k=1}^{m} n_{i,k} = n_i$.

Vzhledem k tomu, že sloupcový průměr je definován rovnicí

$$\bar{y}_i = \frac{1}{n_i} \sum_k y_k n_{i,k}$$

a dále, že platí

$$\sum_k x_i n_{i,k} = x_i \sum_k n_{i,k} = x_i n_i,$$

dostáváme

$$\sum_{i,k} n_{i,k} (y_k - a x_i - b) = 0,$$

$$\sum_{i,k} n_{i,k} (y_k - a x_i - b) x_i = 0,$$

čili

$$\sum_{i,k} n_{i,k} y_k = a \sum_{i,k} n_{i,k} x_i + b \sum_{i,k} n_{i,k},$$

$$\sum_{i,k} n_{i,k} y_k x_i = a \sum_{i,k} n_{i,k} x_i^2 + b \sum_{i,k} n_{i,k} x_i,$$

a jejich řešením vyplývá

$$a = \frac{r \Sigma n_{i,k} x_i y_k - \Sigma n_{i,k} x_i \Sigma n_{i,k} y_k}{r \Sigma n_{i,k} x_i^2 - (\Sigma n_{i,k} x_i)^2},$$

$$b = \frac{\Sigma n_{i,k} y_i \Sigma n_{i,k} x_i^2 - \Sigma n_{i,k} x_i \Sigma n_{i,k} x_i y_k}{r \Sigma n_{i,k} x_i^2 - (\Sigma n_{i,k} x_i)^2},$$

což jsou tytéž rovnice jako (64), takže z nich stejným způsobem dostáváme tytéž rovnice regresních přímek.

(8,1) Mnohonásobná korelace. Od korelačního vztahu mezi dvěma kvantitativními znaky přejdeme nyní ke studiu kolektivní závislosti jednoho znaku na dvou nebo více dalších znacích. Na př. velikost sklizně určité plodiny v jednom roce závisela na vlivu několika činitelů, z nichž nejdůležitějšími jsou množství srážek a tepelné poměry, neboť mají základní význam pro růst rostliny. Těsnost vázanosti mezi velikostí sklizně na jedné straně a množstvím srážek i poměry tepelnými na druhé straně můžeme studovati opět pomocí regresních přímek a korelačního koeficientu, vystihneme-li nějakým způsobem současný vliv několika činitelů. Podobně jakost materiálu se zkouší často podle vztahu mezi hustotou, tvrdostí a tažností. Omezíme se na tři proměnné a odvodíme podobně jako v případě dvou proměnných

koeficient korelace, je-li vztah mezi proměnnými lineární. Odhadujeme-li proměnnou y pomocí x a z, napíšeme lineární vztah

$$y = a_0 + a_1 x + a_2 z$$

čili pro odchylky od průměrů vzhledem ku (75) také

$$\eta = a_1 \xi + a_2 \zeta. \tag{98}$$

Normální rovnice pak jsou

$$\Sigma \xi\eta = a_1 \Sigma \xi^2 + a_2 \Sigma \xi\zeta, \tag{99}$$
$$\Sigma \zeta\eta = a_1 \Sigma \xi\zeta + a_2 \Sigma \zeta^2, \tag{100}$$

z nichž vyplývá řešením

$$a_1 = \frac{\Sigma\zeta^2 \Sigma\xi\eta - \Sigma\xi\zeta \Sigma\zeta\eta}{\Sigma\xi^2 \Sigma\zeta^2 - (\Sigma\xi\zeta)^2}, \quad a_2 = \frac{\Sigma\xi^2 \Sigma\zeta\eta - \Sigma\xi\zeta \Sigma\xi\eta}{\Sigma\xi^2 \Sigma\zeta^2 - (\Sigma\xi\zeta)^2}. \tag{101}$$

a hledaná rovnice (98) pro odhad η je tedy určena. Odvodíme nyní výraz pro čtverec průměrné čtvercové odchylky residuí obdobně jako v případě dvou proměnných.

$$s_{y.xz}^2 = \frac{1}{r} [\Sigma\eta^2 - \Sigma(a_1\xi + a_2\zeta)^2]$$

čili

$$s_{y.xz}^2 = \frac{1}{r} [\Sigma\eta^2 - (a_1^2 \Sigma\xi^2 + a_2^2 \Sigma\zeta^2 + 2a_1 a_2 \Sigma\xi\zeta)].$$

Násobíme-li rovnici (99) koeficientem a_1, rovnici (100) koeficientem a_2 a sečteme, vidíme, že výraz v kulaté závorce je $a_1 \Sigma\xi\eta + a_2 \Sigma\zeta\eta$, takže

$$s_{y.xz}^2 = \frac{1}{r} (\Sigma\eta^2 - a_1 \Sigma\xi\eta - a_2 \Sigma\zeta\eta)$$

a dosadíme-li za hodnoty konstant výraz (101), dostáváme

$$s_{y.xz}^2 = \frac{1}{r} \left\{ \Sigma\eta^2 - \frac{(\Sigma\zeta^2 \Sigma\xi\eta - \Sigma\xi\zeta \Sigma\zeta\eta)\, \Sigma\xi\eta}{\Sigma\xi^2 \Sigma\zeta^2 - (\Sigma\xi\zeta)^2} - \right.$$
$$\left. - \frac{(\Sigma\xi^2 \Sigma\zeta\eta - \Sigma\xi\zeta \Sigma\xi\eta)\, \Sigma\zeta\eta}{\Sigma\xi^2 \Sigma\zeta^2 - (\Sigma\xi\zeta)^2} \right\}$$

čili

$$s_{y.xz}{}^2 = \frac{1}{r}\left\{\Sigma\eta^2 - \frac{\Sigma\zeta^2(\Sigma\xi\eta)^2 - 2\Sigma\xi\zeta\Sigma\zeta\eta\Sigma\xi\eta + \Sigma\xi^2(\Sigma\zeta\eta)^2}{\Sigma\xi^2\Sigma\zeta^2 - (\Sigma\xi\zeta)^2}\right\},$$

$$s_{y.xz}{}^2 = \frac{\Sigma\eta^2}{r}\left\{1 - \frac{\Sigma\zeta^2(\Sigma\xi\eta)^2 - 2\Sigma\xi\zeta\Sigma\zeta\eta\Sigma\xi\eta + \Sigma\xi^2(\Sigma\zeta\eta)^2}{[\Sigma\xi^2\Sigma\zeta^2 - (\Sigma\xi\zeta)^2]\,\Sigma\eta^2}\right\},$$

což napíšeme

$$s_{y.xz}{}^2 = \sigma_y{}^2\,(1 - r_{y.xz}{}^2), \tag{102}$$

kde jsme zavedli pro zlomek ve velké závorce označení $r_{y.xz}{}^2$; dělíme-li v něm čitatele i jmenovatele součinem $\Sigma\xi^2\Sigma\eta^2\Sigma\zeta^2$, dostáváme

$$r_{y.xz}{}^2 = \frac{\dfrac{(\Sigma\xi\eta)^2}{\Sigma\xi^2\Sigma\eta^2} - \dfrac{2\Sigma\xi\zeta\Sigma\zeta\eta\Sigma\xi\eta}{\Sigma\xi^2\Sigma\eta^2\Sigma\zeta^2} + \dfrac{(\Sigma\zeta\eta^2)}{\Sigma\zeta^2\Sigma\eta^2}}{1 - \dfrac{(\Sigma\xi\zeta)^2}{\Sigma\xi^2\Sigma\zeta^2}}.$$

Zavedeme-li pak podle rovnice (77) příslušné symboly koeficientů korelace mezi dvěma znaky, můžeme psáti poslední rovnici

$$r_{y.xz}{}^2 = \frac{r_{xy}{}^2 - 2r_{xy}r_{yz}r_{xz} + r_{yz}{}^2}{1 - r_{yz}{}^2}. \tag{103}$$

Tak je vyjádřen koeficient mnohonásobné korelace pro tři proměnné mezi znakem y a dvěma znaky x a z. Můžeme si představiti, že tečky stereogramu rozdělení četností, který bychom si sestrojili v trojrozměrném, prostoru jsou rozptýleny kolem roviny regrese. Analogicky k rovnici (103) lze snadno napsati příslušné výrazy pro $r_{z.xy}{}^2$ a $r_{x.yz}{}^2$.

Na pořadí indexů jednotlivých koeficientů korelace nezáleží, takže $r_{yz} = r_{zy}$ a $r_{xz} = r_{zx}$ a tudíž také nezáleží na pořadí indexů za tečkou ve (103), takže $r_{y.xz}{}^2 = r_{y.zx}{}^2$ a obdobně $r_{z.xy}{}^2 = r_{z.yx}{}^2$, $r_{x.yz}{}^2 = r_{x.zy}{}^2$.

Úvahy provedené zde pro tři proměnné lze rozšířiti na libovolný počet proměnných [1].

(8,1,1) Příklad 1. Vyjádřete rovinu regrese hodnot proměnné z vzhledem k proměnným x a y pomocí příslušných směrodatných odchylek a koeficientů korelace. Zvolíme-li za počátek souřadnic průměr celého rozdělení četností v trojrozměrném prostoru, bude rovnice této roviny analogicky ku (98)

$$\zeta = b_1\xi + b_2\eta,$$

kde koeficienty b_1 a b_2 jsou vyjádřeny analogickými rovnicemi ku (101)

$$b_1 = \frac{\Sigma\eta^2\Sigma\xi\zeta - \Sigma\xi\eta\Sigma\eta\zeta}{\Sigma\xi^2\Sigma\eta^2 - (\Sigma\xi\eta)^2}, \qquad b_2 = \frac{\Sigma\xi^2\Sigma\eta\zeta - \Sigma\xi\eta\Sigma\xi\zeta}{\Sigma\xi^2\Sigma\eta^2 - (\Sigma\xi\eta)^2}$$

a vzhledem k rovnici (78) a obdobným pro ostatní dvojice proměnných vyplývá po malé úpravě

$$b_1 = \frac{\sigma_z\,(r_{xz} - r_{yz}r_{xy})}{\sigma_x\,(1 - r_{xy}^2)}, \qquad b_2 = \frac{\sigma_z\,(r_{yz} - r_{xy}r_{xz})}{\sigma_y\,(1 - r_{xy}^2)}.$$

Čtverec průměrné čtvercové odchylky residuí můžeme pak napsati pomocí determinantu ve tvaru

$$s_{z.xy}^2 = \frac{\sigma_z}{1 - r_{xy}^2} \cdot \begin{vmatrix} 1 & r_{yz} & r_{xz} \\ r_{yz} & 1 & r_{xy} \\ r_{xz} & r_{xy} & 1 \end{vmatrix}. \qquad (104)$$

Rozvedením determinantu podle prvního řádku nebo sloupce je možno se přesvědčit o totožnosti tohoto vyjádření s tím, které odpovídá rovnicím (102) a (103).

Příklad 2. Jakost určitého materiálu byla charakterisována třemi znaky x, y, z, které byly pozorovány na souboru třiceti prvků; pozorované hodnoty v příslušných jednotkách každé proměnné, v nichž byla měřena, jsou sestaveny v tab. 21.

Tab. 21.

i	x	y	z	i	x	y	z
1	2,7	71,4	35,4	16	2,5	55,7	28,8
2	2,6	53,4	31,3	17	2,7	70,5	34,0
3	2,7	82,5	32,2	18	2,9	87,5	34,5
4	2,6	67,3	33,4	19	2,6	50,7	29,9
5	2,5	69,5	37,7	20	2,4	59,5	29,8
6	2,7	73,0	34,9	21	2,6	71,3	29,3
7	2,6	55,7	24,7	22	2,7	76,5	31,4
8	2,8	85,8	34,7	23	2,6	69,2	31,7
9	2,8	95,4	38,0	24	2,8	83,7	36,8
10	2,5	51,1	25,7	25	2,9	94,7	41,6
11	2,6	74,2	25,8	26	2,7	70,2	30,5
12	2,6	77,6	28,0	27	2,6	80,4	29,7
13	2,5	64,1	25,8	28	2,7	76,7	32,6
14	2,6	53,7	23,7	29	2,6	78,0	29,2
15	2,7	82,2	32,4	30	2,8	79,3	36,7

Jest najíti základní charakteristiky jednotlivých řad a vztahů mezi nimi, sestrojiti a znázorniti rovinu regrese pro odhad hodnot proměnné z vzhledem ku x a y.

Známým postupem zjistíme

$$\bar{x} = 2{,}65, \quad \sigma_x = 0{,}18, \quad r_{xy} = 0{,}60,$$
$$\bar{y} = 72{,}03, \quad \sigma_y = 12{,}16, \quad r_{yz} = 0{,}67,$$
$$\bar{z} = 31{,}67, \quad \sigma_z = 4{,}23, \quad r_{xz} = 0.59$$

Rovnice přímek regrese

$$\zeta = r_{yz}\frac{\sigma_z}{\sigma_y}\eta = 0{,}23\eta, \quad \zeta = r_{xz}\frac{\sigma_z}{\sigma_x}\xi = 13{,}86\xi,$$

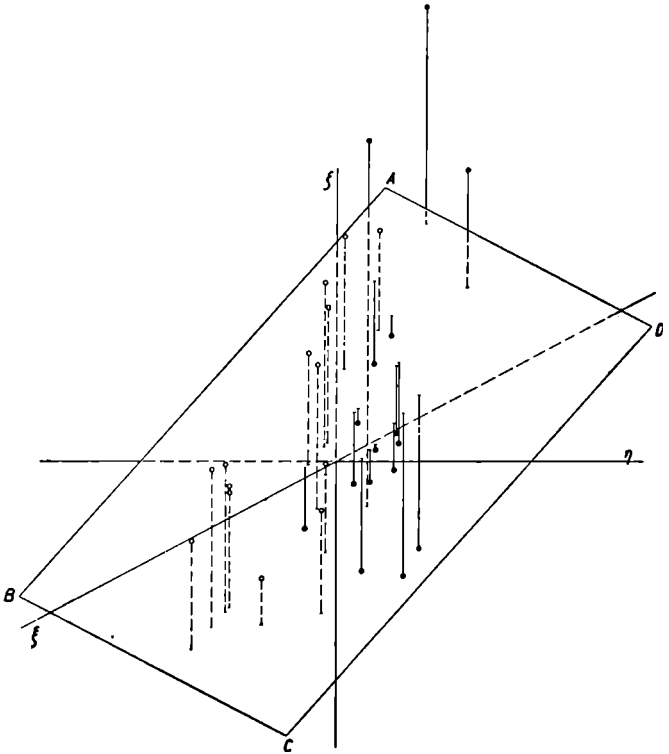
$$s_{zy} = \sigma_z\sqrt{1 - r_{yz}^2} = 3{,}13, \quad s_{zx} = \sigma_z\sqrt{1 - r_{xz}^2} = 3{,}43.$$

Rovnice regresní roviny

$$\zeta = b_1\xi + b_2\eta,$$
$$b_1 = \frac{\sigma_z\,(r_{xz} - r_{yz} + r_{xy})}{\sigma_x\,(1 - r_{xy}^2)} = 6{,}977,$$
$$b_2 = 0{,}174,$$

bude pak $\zeta = 6{,}977\xi + 0{,}174\eta$

a $s_{z.xy} = 8{,}63, \qquad r_{z.xy}{}^2 = 0{,}5.$



Obr. 18. Regresní rovina.

Grafické znázornění regresní roviny i celého roje teček je provedeno v obr. 18. Průměrná čtvercová odchylka residuí kolem roviny je víc než dvojnásobkem průměrných čtvercových odchylek kolem regresních přímek.

Úlohy: 1. Proveďte rozbor případu, který nastane a) jestliže v rovnici (103) je

$$r_{xy} = r_{yz} = 0,$$

takže

$$r_{y.xz}^2 = \frac{2r_{xy}^2}{1 + r_{xz}} = 0$$

b) je-li

$$r_{xy} = r_{y.xz} = 0,$$

takže

$$r_{yz} = 0.$$

c) je-li

$$r_{xz} = 1$$

d) $$r_{xy} = r_{xz} = r_{yz} = r_{y.xz},$$

takže všechny koeficienty korelace se rovnají $-\frac{1}{2}$.

2. Najděte $r_{y.xz}$, je-li $r_{xy} = r_{xi} = r_{yz} = 0{,}999$.

3. Proveďte rozbor případu, který nastane, jestliže v rovnici (102) je $r_{xy} = r_{yz} = 1$.

(8,2) Dílčí korelace. Představme si nyní, že máme zjistit těsnost vztahu mezi velikostí sklizně y a tepelnými poměry x, k čemuž vypočítáme koeficient korelace r_{xy}. Nahlédneme snadno, že nebude představovati výstižně a bezpečně tento vztah, takže nebudeme moci odhadovati vzrůst velikosti sklizně odpovídající určitému vzrůstu průměrné teploty, poněvadž teplota je v těsném vztahu s některými dalšími důležitými činiteli, jako je na př. množství srážek z. V takovém případě musíme užíti k řešení úkolu zvláštního postupu, kterým stanovíme korelaci mezi velikostí sklizně a tepelnými poměry, když zůstává množství srážek konstantní. Je to t. zv. dílčí korelace, která je určena koeficientem dílčí korelace $r_{xy.z}$ mezi dvěma proměnnými při určité stálé hodnotě třetí proměnné. Tento koeficient může býti stálý, nebo se může měniti, když třetí proměnná nabývá různých hodnot. Abychom dospěli k jednoduchému odvození hodnoty koeficientu dílčí korelace, připomeneme si, že jednoduchý koeficient korelace r_{xy} mezi dvěma proměnnými jsme dostali také jako

geometrický průměr koeficientů regrese v rovnicích přímek regresních

$$\eta = b_{21}\xi \quad \text{a} \quad \xi = b_{12}\eta.$$

Uvažujeme tedy tři proměnné, které jsou v lineárním vztahu vyjádřeném v odchylkách od průměrů rovnicí

$$\eta = a_1\xi + a_2\zeta, \tag{105}$$

kde koeficient a_1 je dán podle (101) výrazem

$$a_1 = \frac{\Sigma\zeta^2\Sigma\xi\eta - \Sigma\xi\zeta\Sigma\zeta\eta}{\Sigma\xi^2\Sigma\zeta^2 - (\Sigma\xi\zeta)^2}.$$

Rovnice lineárního vztahu pro odhad ξ pomocí η a ζ je analogicky

$$\xi = c_1\eta + c_2\zeta, \tag{106}$$

kde koeficienty c_1 a c_2 dostaneme opět řešením příslušných normálních rovnic

$$\Sigma\xi\eta = c_1\Sigma\eta^2 + c_2\Sigma\zeta\eta,$$
$$\Sigma\xi\zeta = c_1\Sigma\eta\zeta + c_2\Sigma\zeta^2,$$

odkud dostaneme pro c_1 výsledek

$$c_1 = \frac{\Sigma\zeta^2\Sigma\xi\eta - \Sigma\zeta\eta\Sigma\xi\zeta}{\Sigma\eta^2\Sigma\zeta^2 - (\Sigma\eta\zeta)^2}$$

a utvoříme-li geometrický průměr obou koeficientů a_1 a c_1, dostaneme

$$\sqrt{a_1c_1} = \left\{\frac{(\Sigma\zeta^2)^2(\Sigma\xi\eta)^2 - 2\Sigma\zeta^2\Sigma\xi\eta\Sigma\zeta\eta\Sigma\xi\zeta + (\Sigma\xi\zeta)^2(\Sigma\zeta\eta)^2}{[\Sigma\xi^2\Sigma\zeta^2 - (\Sigma\xi\zeta)^2][\Sigma\zeta^2\Sigma\eta^2 - (\Sigma\eta\zeta)^2]}\right\}^{\frac{1}{2}},$$

což můžeme přepsati pomocí symbolů jednoduchých koeficientů korelace mezi dvěma proměnnými

$$\sqrt{a_1c_1} = \sqrt{\frac{r_{xy}^2 - 2r_{xy}r_{xz}r_{yz} + r_{xz}^2r_{yz}^2}{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

Koeficient dílčí korelace mezi x a y je definován jako tento geometrický průměr koeficientů regrese a_1 a c_1, takže píšeme

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)\,(1 - r_{yz}^2)}}. \qquad (107)$$

Z odvození je patrno, že hodnot koeficientů a_2 a c_2 při ζ v rovnicích (105) a (106) nebylo použito, což znamená, že se ponechává ζ konstantní a právě proto, že byla získána hodnota koeficientu korelace mezi dvěma proměnnými při konstantní hodnotě třetí proměnné, nazývá se koeficientem dílčí korelace. Při užívání tohoto koeficientu musíme se přesvědčit, jsou-li všechny dvojice proměnných v lineárním vztahu vzájemném. Výpočet pak lze podstatně usnadnit tabulkami jako na př. [3]. Rozšíření na více proměnných je možno provésti zcela obecně [1].

(8,2,1) Příklad 1. Předpokládejme, že byl zjištěn jednoduchý koeficient korelace mezi velikostí sklizně y a teplotou x hodnotou $r_{xy} = +\,0{,}62$, mezi velikostí sklizně y a množstvím srážek z hodnotou $r_{yz} = +\,0{,}80$, mezi teplotou a množstvím srážek $r_{xz} = +\,0{,}75$. Jaký bude koeficient dílčí korelace mezi velikostí sklizně a teplotou?

Podle rovnice (107) bude na př. pomocí tabulek [3] $r_{xy.z} = \frac{0{,}62 - 0{,}60}{0{,}397} = +\,0{,}05$, takže skutečný vliv teploty při stálém množství srážek prakticky mizí.

Příklad 2. Jest stanoviti koeficient dílčí korelace mezi z a x, jakož i mezi z a y pro materiál tabulky 17.

$$r_{xz.y} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{(1 - r_{xy}^2)\,(1 - r_{yz}^2)}} = \frac{0{,}587 - 0{,}602 \,.\, 0{,}671}{0{,}594} = 0{,}31,$$

$$r_{yz.x} = \frac{r_{yz} - r_{xz}r_{xy}}{\sqrt{(1 - r_{xy}^2)\,(1 - r_{xz}^2)}} = \frac{0{,}671 - 0{,}354}{0{,}646} = 0{,}49.$$

Příklad 3. Dokažte, že mezi koeficienty korelace platí vztah

$$(1 - r_{xy.z}^2)\,(1 - r_{xy}^2) = (1 - r_{y.xz}^2).$$

Z rovnice (107) a (103) dosadíme a dostaneme

$$(1 - r_{yz}^2) - \frac{(r_{xy} - r_{xz}r_{yz})^2}{1 - r_{xz}^2} =$$

$$= \frac{(1 - r_{xz}^2) - (r_{xy}^2 - 2r_{xy}r_{yz}r_{xz} + r_{yz}^2)}{1 - r_{xz}^2},$$

což je identita, neboť pravou stranu můžeme uvésti na tvar

$$\frac{1 - r_{xz}^2 - r_{yz}^2 + r_{xz}^2 r_{yz}^2 - (r_{xy} - r_{xz}r_{yz})^2}{1 - r_{xz}^2} =$$

$$= \frac{(1 - r_{xz}^2)(1 - r_{yz}^2)}{1 - r_{xz}^2} - \frac{(r_{xy} - r_{xz}r_{yz})^2}{1 - r_{xz}^2}.$$

Úlohy: 1. Proveďte rozbor případu, který nastane, když ve formuli (107)

a) $r_{xy.z}^2 = 1$ b) $r_{xz}^2 = 1$ c) $r_{xy}^2 = 1$

2. Dokažte, že pro čtyři proměnné platí

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\,r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}},$$

kde proměnné jsou označeny číslicemi.

(8,3) Korelační poměr (vztah nelineární). Předpokladem upotřebitelnosti koeficientu korelace je lineární regrese. Pro případy nelineární korelace je užitečnou mírou těsnosti vztahu korelační poměr y vzhledem ku x, který označujeme η_{yx} a je tedy obecnější mírou korelace. V případech, kdy je pochybno, zda je vztah mezi proměnnými skutečně lineární, je výpočet korelačního poměru nutnou součástí korelační analysy. Obdobně jako čtverec průměrné čtvercové odchylky residuí s_{xy}^2 (66) definujeme rozptyl kolem sloupcových průměrů s_y^2; odchylky se neměří od přímky odhadu, tedy od přímky regrese, nýbrž od příslušných sloupcových průměrů.

Podle toho na př. od každé hodnoty y prvního sloupce odečteme průměr prvního sloupce a tak to provedeme v každém

sloupci. Průměr čtverců těchto odchylek od příslušných průměrů je právě rozptyl kolem sloupcových průměrů s_y^2. Podle toho bude

$$s_y^2 = \frac{1}{r}\{\Sigma_1(y_i - \bar{y}_1)^2 + \Sigma_2(y_i - \bar{y}_2)^2 + \ldots + \Sigma_l(y_i - \bar{y}_l)^2\}. \tag{108}$$

je-li l celkový počet sloupců. Budeme-li označovati součet hodnot y v i-tém sloupci s_i, pak průměr $\bar{y}_i = \frac{s_i}{n_i}$, neboť marginální četnost sloupce je n_i, a rovnici (108) můžeme psáti

$$s_y^2 = \frac{1}{r}\left\{\Sigma_1\left(y_i - \frac{s_1}{n_1}\right)^2 + \Sigma_2\left(y_i - \frac{s_2}{n_2}\right)^2 + \ldots + \Sigma_l\left(y_i - \frac{s_l}{n_l}\right)^2\right\}.$$

Vzhledem k tomu, že

$$\Sigma_1\left(y_i - \frac{s_1}{n_1}\right)^2 = \Sigma_1 y_i^2 - 2\frac{s_1}{n_1}\Sigma_1 y_i + \frac{s_1^2}{n_1} = \Sigma_1 y_i^2 - \frac{s_1^2}{n_1}$$

a obdobně je tomu pro ostatní součty velké závorky, můžeme psáti

$$s_y^2 = \frac{1}{r}\left\{\Sigma_1 y_i^2 - \frac{s_1^2}{n_1} + \Sigma_2 y_i^2 - \frac{s_2^2}{n_2} + \ldots + \Sigma_l y_i^2 - \frac{s_l^2}{n_l}\right\}$$

čili

$$s_y^2 = \frac{1}{r}\Sigma y^2 - \frac{1}{r}\sum_i \frac{s_i^2}{n_i},$$

kde první součet se vztahuje na všechny hodnoty y souboru a druhý součet na všechny sloupce, tedy $i = 1, 2, \ldots, l$. Odečteme-li a přičteme nyní $\bar{y}^2$, dostaneme

$$s_y^2 = \frac{\Sigma y^2}{r} - \bar{y}^2 - \left[\frac{1}{r}\sum_i \frac{s_i^2}{n_i} - \bar{y}^2\right] = \sigma_y^2 - \left[\frac{1}{r}\sum_i \frac{s_i^2}{n_i} - \bar{y}^2\right]$$

$$s_y^2 = \sigma_y^2\left\{1 - \frac{\frac{1}{r}\sum_i \frac{s_i^2}{n_i} - \bar{y}^2}{\sigma_y^2}\right\},$$

$$s_y^2 = \sigma_y^2(1 - \eta_{yx}^2) \tag{109}$$

čili korelační poměr η_{yx}^2 z této rovnice bude

$$\eta_{yx}^2 = 1 - \frac{s_y^2}{\sigma_y^2}. \tag{110}$$

Zavedli jsme tedy pro korelační poměr výraz

$$\eta_{yx}^2 = \frac{\frac{1}{r}\sum_i \frac{s_i^2}{n_i} - \bar{y}^2}{\sigma_y^2}. \tag{111}$$

V souboru, kde je lineární regrese, bude $s_{xy}^2 = s_y^2$, takže porovnáním rovnic (83) a (110) vidíme, že $\eta_{yx}^2 = r_{xy}^2$. Z rovnice (110) je patrno, že $\eta_{yx}^2 \leqq 1$, neboť zlomek je tam vždy veličina kladná a rovnost může nastati jen tehdy, když všechny hodnoty každého sloupce se rovnají jeho průměru, neboť pak je $s_y^2 = 0$. Není-li s_{xy}^2 rovno s_y^2, musí býti větší, neboť průměr čtverců odchylek hodnot proměnné v jednom sloupci od nějaké hodnoty je nejmenší pro odchylky od průměru tohoto sloupce.

Z toho pak vyplývá, že

$$\eta_{yx}^2 \geqq r_{xy}^2.$$

K posouzení lineárnosti regrese se používá rozdílu $\eta_{yx}^2 - r_{xy}^2$, při čemž je třeba přihlížeti k variacím náhodného výběru, o jejichž testování se může čtenář dovědět bližší v [1].

Poněvadž platí také mezi rozptyly vztah

$$\sigma_{\bar{y}_i}^2 = \sigma_y^2 - s_y^2,$$

který si ověříme dosazením příslušných výrazů za

$$\sigma_y^2 = \frac{1}{r}\sum_{i=1}^{l}\sum_i (y - \bar{y})^2, \qquad s_y^2 = \frac{1}{r}\sum_{i=1}^{l}\sum_i (y - \bar{y}_i)^2$$

a rozptyl sloupcových průměrů

$$\sigma_{\bar{y}_i}^2 = \frac{\sum_{i=1}^{l} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{l} n_i},$$

neboť po dosazení a vynásobení rozsahem r dostáváme součet $\sum_{i=1}^{l}$ identit

$$\Sigma_i(y - \bar{y})^2 = \Sigma_i(y - \bar{y}_i)^2 + n_i(\bar{y}_i - \bar{y})^2,$$

jež vyplývají z rovnice

$$(y - \bar{y})^2 = [(y - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2$$

vzhledem k tomu, že

$$2\,(\bar{y}_i - \bar{y})\,\Sigma_i(y - \bar{y}_i) = 0.$$

Platí tudíž také rovnice

$$\eta_{yx}^2 = \frac{\sigma_{y_i}^2}{\sigma_y^2}, \tag{112}$$

takže korelační poměr je také poměr rozptylu průměrů sloupcových k rozptylu proměnné y v celém souboru.

Co bylo řečeno o η_{yx}^2, platí obdobně také o $\eta_{xy}^2 = 1 - \frac{s_x^2}{\sigma_x^2}$, a počítá se podle formule

$$\eta_{xy}^2 = \frac{\frac{1}{r}\sum_k \frac{s_k^2}{n_k} - \bar{x}^2}{\sigma_x^2}. \tag{113}$$

(8,3,1) Příklad. Proveďme výpočet korelačního poměru pro soubor v tab. 14. Použijeme k tomu sloupců $\frac{s_i^2}{n}$ a $\frac{s_k^2}{n_k}$, počítaných pro proměnné w, v, takže dostaneme

$$\frac{\Sigma \frac{s_i^2}{n_i}}{r} = \frac{147{,}94}{156} = 0{,}9483$$

a tedy podle rovnice (111)

$$\eta_{yx}^2 = \frac{1}{\sigma_w^2}\left\{\frac{\Sigma \frac{s_i^2}{n_i}}{r} - \bar{w}^2\right\},$$

bude

$$\eta_{yz} = 0{,}705.$$

Podobně pak

$$\frac{\Sigma \frac{s_k^2}{n_k}}{r} = \frac{110{,}52}{156} = 0{,}7085, \qquad \eta_{xy}^2 = \frac{1}{\sigma_v^2}\left\{\frac{\Sigma \frac{s_k^2}{n_k}}{r} - \bar{v}^2\right\}$$

čili

$$\eta_{xy} = 0{,}694.$$

(8,4) Meze užití koeficientu korelace. Je třeba upozorniti čtenáře, že nelze přikládati příliš mnoho důležitosti koeficientům korelace počítaným z malého počtu prvků. Má-li se použíti určitého číselného výsledku pro koeficient korelace jako základu pro odůvodňování obecně platné, musí býti vypočítán ze souboru dostatečného rozsahu. Ovšem i řádně zjištěný vysoký stupeň korelace nemusí býti průkazem, že vlastnost popsaná jedním znakem je příčinou druhé vlastnosti, popsané druhým znakem.

Dříve než označíme jeden z korelovaných znaků za příčinu a druhý za následek, je důležito zkoumati, zda obě množiny čísel popisující tyto dva znaky nemohou býti výsledkem nějakého činitele třetího. Posouzení významu velikosti koeficientu korelace je přirozeně pro jednotlivé obory, v nichž se ho užívá, velmi různé. Tam, kde můžeme předpokládati, že obě řady čísel jsou vzájemně vázány lineárním vztahem, jako je tomu v mnohých aplikacích technických, nebudeme koeficient korelace ve výši 0,9 hodnotiti příliš vysoko, kdežto při rozborech populačně statistických nebo hospodářských dat, pro něž nemůže existovati přesně platný teoretický vztah, může býti koeficient korelace 0,8 hodnocen v některých případech za velmi značný. Nemůžeme se spokojit jen s korelačním výpočtem, nýbrž množství úvah přípravných spočívajících ve vhodném položení otázky a racionelním rozčlenění může poskytovati prakticky vědeckou bezpečnost o příčinných vztazích. Teorie korelace má v bádání příčin-

ném nepřekročitelnou hranici v tom, že musí pracovat jako součást statistiky s empirickým materiálem, který je jí dán a který si nemůže jinak opatřit. Když pak začíná výpočet, byla již hlavní práce vykonána a výsledku vlastně bylo již dosaženo. Bylo provedeno při užívání teorie korelace v minulosti mnoho chyb a nejvíce jich spadá do oblasti, kterou nazýváme logika korelace. Sem patří především zkoumání možnosti nějakého vztahu, způsob kladení otázky a kritika kladení otázky. Musíme míti na paměti, že příčinné bádání statistiky nespočívá na jejím matematickém, nýbrž hlavně na jejím logickém a noetickém základu, jehož vady a nedostatky s ním tudíž sdílí. Koeficient korelace je jakýmsi indexem vztahu, nikoliv důkazem příčinné vázanosti. Počítá se, jako jiné statistické charakteristiky, za účelem osvětlení výkladu a rozboru velikých množství pozorovaných dat. Tento výklad musí býti v souhlasu se zdravou logickou analysou. Prakse každého oboru si ustálí obyčejně nějaké rozdělení celé stupnice koeficientu korelace od 0 do 1, takže na př. usuzuje, že koeficient korelace

1. menší než 0,3 naznačuje nízký stupeň těsnosti vztahu a nespoléhá příliš na významnost jeho, je-li zvláště rozsah souboru malý.
2. $0{,}3 \leqq r_{xy} < 0{,}5$ naznačuje mírný stupeň těsnosti vztahu, je-li jeho pravděpodobná chyba malá.
3. $0{,}5 \leqq r_{xy} < 0{,}7$ ukazuje na význačnou těsnost vztahu.
4. $0{,}7 \leqq r_{xy} < 0{,}9$ je ukazatelem vysokého stupně těsnosti.
5. $0{,}9 \leqq r_{xy}$ značí velmi těsný vztah, čili velmi vysoký stupeň vázanosti mezi proměnnými.

Je však velmi důležito míti stále na paměti, že interpretace významnosti nezávisí jen na velikosti koeficientu, nýbrž také na rozsahu pozorovaného souboru. Je-li koeficient nízký nebo mírný, jsou jeho náhodně výběrové odchylky takové, že jej činí nespolehlivým a pochybné významnosti, je-li rozsah

výběru malý. Spolehlivost takových výsledků se může zvýšiti, lze-li opakovati pozorování na mnoha takových malých výběrech.

Na konec musíme zvláště zdůrazniti předpoklad našich dosavadních vývodů, že hodnoty proměnné byly měřeny přesně. Nepřihlíželi jsme tedy k chybám měření, t. j. k odchylkám zjištěných hodnot od skutečných. Tato okolnost má zvláštní význam při uvažování významu charakteristik s hlediska odchylek výběrových.

(9) Koeficient korelace s hlediska teorie náhodného výběru.

(9,1) Hypotésa nulová. Výběrové charakteristiky mají svá rozdělení četností, která nám pomáhají odhadovati jejich odchylky od příslušných parametrů v základním souboru, jež se vyskytují s určitými pravděpodobnostmi. Testujeme tak jejich významnost. Zcela obdobně si představíme, že charakteristiky, které se při dvojrozměrném třídění k dřívějším připojily, mají také svá rozdělení četností, takže úvahy provedené v první části budou zde míti svoji obdobu.

Začneme s nejjednodušším a velmi obvyklým testem, kterým zjišťujeme, je-li nějaký pozorovaný koeficient korelace významně větší než nula. Je to případ, v němž máme najíti pravděpodobnost, že taková hodnota r_{xy} jako je pozorovaná v uvažovaném náhodném výběru, by se mohla vyskytnouti v náhodném výběru z nějakého základního souboru, v němž znaky x a y nejsou ve vztahu, čili koeficient korelace $r(x, y) = 0$. Bylo dovozeno, že pro výběry z takového základního souboru má rozdělení četností hodnot r_{xy} směrodatnou odchylku

$$\sigma(r_{xy}) = \frac{1}{\sqrt{r - 1}}, \qquad (114)$$

kde r ve jmenovateli značí počet dvojic hodnot x a y, čili počet prvků výběru. Není-li výběr příliš malý, je rozdělení r_{xy} dosti blízké normálnímu, takže je oprávněno a postačí užíti kriteria, že nějaká hodnota r_{xy}, větší než dvojnásobná směro-

datná odchylka (114) je nad 5% hladinou významnosti. V takových případech tedy stačí najíti $r_{xy}\sqrt{r-1}$ a užíti tabulky Laplaceova integrálu, abychom určili pravděpodobnost P, že pozorovaná hodnota r_{xy} koeficientu korelace se mohla vyskytnouti v náhodném výběru z nějakého základního souboru, v němž $r(x, y) = 0$. Uvedený výraz $r_{xy}\sqrt{r-1}$ je totiž $\frac{r_{xy}}{\sigma(r_{x,y})}$, což tu znamená odchylku od průměru (který je v bodě nula), vyjádřenou ve směrodatné odchylce jako jednotce. Při dosti velikém rozsahu výběru je vyhovujícím přiblížením směrodatné odchylky $\frac{1}{\sqrt{r}}$.

(9,2) Malé výběry. Pro malé výběry se rozdělení četností hodnot r_{xy} nepřibližuje dosti těsně normálnímu, takže pak předcházející postup testování není oprávněn. Musíme potom užíti především výstižnějšího výrazu pro směrodatnou odchylku koeficientu korelace

$$\sigma^2(r_{xy}) = \frac{[1 - r^2(x, y)]^2}{r - 1}. \tag{115}$$

Ve velkých výběrech a při nevelké těsnosti vztahu má koeficient korelace z výběru o r dvojicích normální rozdělení kolem hodnoty $r(x, y)$ v základním souboru, se směrodatnou odchylkou (115). Tento výraz však obsahuje parametr $r(x, y)$, který neznáme zpravidla a nahrazujeme jej pomocí pozorované hodnoty koeficientu korelace

$$\sigma_{r_{xy}} = \frac{1 - r_{xy}^2}{\sqrt{r-1}}. \tag{116}$$

Při malých výběrech se může hodnota r_{xy} velmi lišiti od $r(x, y)$, takže v čitateli se můžeme dopustit značné chyby a nad to rozdělení hodnot koeficientu korelace se velmi liší od normálního. Aby bylo umožněno provedení testu také v případech, kde rozsah výběrů je menší než sto, které se

v některých oborech aplikace často vyskytují, používá se t-testu. Zvolí-li se totiž za hodnotu t výraz

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n}, \tag{117}$$

kde $n = r - 2$ je počet stupňů volnosti, lze dokázati, že rozdělení četnosti hodnot t takto počítaných se shoduje s rozdělením tabulky hodnot t (tab. 5).

Dva stupně volnosti byly ztraceny výpočtem dvou charakteristik výběrových zahrnutých ve výrazu pro koeficient korelace. Pomocí tohoto výrazu (117) testujeme tedy významnost pozorovaného koeficientu korelace tak, že chceme zjistiti pravděpodobnost, že taková hodnota jeho se může vyskytnouti v náhodném výběru ze základního souboru, v němž není vztahu mezi uvažovanými znaky.

(9,3) Korelační transformace z'. Kromě potřeby testovati významnost nějakého koeficientu korelace, abychom zjistili zda můžeme usuzovati na existenci vztahu vůbec, setkáváme se ještě s úkoly dalšími. Bývá třeba testovati, liší-li se pozorovaný koeficient korelace významně od nějaké teoretické hodnoty, nebo zda se dva pozorované koeficienty korelace liší od sebe významně. Jindy docílíme několika nezávislých odhadů určitého koeficientu korelace a máme je kombinovati v jeden zdokonalený odhad, pro který je třeba provésti některý z obou testů uvedených v předcházející větě. Tyto úkoly bychom mohli řešiti analogicky jako jsme testovali významnost r_{xy}, ale obtíže, které jsme tam naznačili, by zde vystupovaly ve zvýšené míře. Byla proto zavedena účelná transformace koeficientu korelace rovnicí

$$z' = \tfrac{1}{2} \{\lg (1 + r_{xy}) - \lg (1 - r_{xy})\} = \\ = r_{xy} + \tfrac{1}{3} r_{xy}^3 + \tfrac{1}{5} r_{xy}^5 + \dots \tag{118}$$

Takto zavedená nová forma charakteristiky má přibližně normální rozdělení četností i pro malé výběry a pomocí ní lze provésti uvedené testy bez obtíží. Směrodatná odchylka tohoto normálního rozdělení četností je

$$\sigma_{z'} = \frac{1}{\sqrt{r-3}}, \tag{119}$$

což je výraz jednoduchý a nezávislý na z'. Jeho veliká výhoda proti (115) je v tom, že je nezávislý na koeficientu korelace v základním souboru, z něhož byl výběr vzat. Probíhá-li r_{xy} hodnoty od 0 do 1, projde z' od 0 do $+\infty$; pro malé hodnoty r_{xy} je z' skoro rovno r_{xy} podle (118), ale když se r_{xy} blíží 1, roste z' nade všechny meze. Je tudíž význam této transformace také v tom, že dává otevřenější stupnici hodnot z', když těsnost vztahu je vysoká. Můžeme shrnouti výhody charakteristiky z' ve tři body. 1. Její směrodatná odchylka je jednoduchý výraz nezávislý na $r(x, y)$. 2. Rozdělení četnosti z' je velmi blízké normálnímu i pro výběry malého rozsahu a s rostoucím rozsahem se normálnímu rozdělení brzo a těsně přimyká ať je hodnota r_{xy} jakákoliv. 3. Kdežto forma rozdělení četností r_{xy} se rychle mění pro měnící se $r(x, y)$, je tvar rozdělení četností z' přibližně konstantní.

K testování významnosti rozdílu mezi dvěma pozorovanými hodnotami koeficientu korelace ${}_1r_{xy}$ a ${}_2r_{xy}$ stanovíme diferenci $z'_1 - z'_2$ pomocí rovnice (118), podle níž bude

$$z'_1 = \tfrac{1}{2}\,[\lg(1 + {}_1r_{xy}) - \lg(1 - {}_1r_{xy})],$$
$$z'_2 = \tfrac{1}{2}\,[\lg(1 + {}_2r_{xy}) - \lg(1 - {}_2r_{xy})],$$

a použijeme směrodatné odchylky

$$\sigma_{z'_1 - z'_2} = \sqrt{\frac{1}{r_1 - 3} + \frac{1}{r_2 - 3}},$$

kterou jsme utvořili podle rovnice [I, (67′)]. Příslušnou pravděpodobnost pak najdeme pro $\dfrac{z'_1 - z'_2}{\sigma_{z'_1 - z'_2}}$ v tabulce Laplaceova integrálu.

(9,3,I) Příklad 1. Budeme testovati významnost koeficientů korelace vypočítaných pro tři soubory tab. 10, v příkladu 1 na str. 112.

Pro soubor

$$\text{č. 1 je } r_{xy} = +0{,}03, \qquad t = \frac{0{,}03\sqrt{11}}{\sqrt{1-0{,}03^2}} = 0{,}10,$$

$$\text{č. 2 je } r_{xy} = +0{,}97, \qquad t = \frac{0{,}97\sqrt{11}}{\sqrt{1-0{,}97^2}} = 13{,}25,$$

$$\text{č. 3 je } r_{xy} = -0{,}98, \qquad t = \frac{0{,}98\sqrt{11}}{\sqrt{1-0{,}98^2}} = 16{,}35.$$

Z tab. 5 seznáme, že pro $n = 11$ je hodnota t na pětiprocentní hranici významnosti mnohem větší než v případě souboru č. 1, takže koeficient korelace je zcela nevýznamný, kdežto v ostatních dvou případech jsou to hodnoty velmi významné.

Příklad 2. Testujme významnost koeficientu korelace, který jsme vypočítali pro soubor roztříděný v tab. 11 v příkladu 2, str. 114. Poněvadž rozsah souboru je $r = 200$, můžeme užíti prvního i druhého způsobu. Podle prvního způsobu bude $r_{xy}\sqrt{199} = 0{,}60\sqrt{199} = 8{,}46$ a pro tuto hodnotu se přesvědčíme v tabulce Laplaceova integrálu, že $P = (1 - \alpha(t))$ je hodnota velmi malá a tudíž významnost vysoká. Podle t-testu dostáváme $t = 10{,}55$ a z tabulky 5 hodnot t pro $n = \infty$ vidíme rovněž vysokou významnost.

Příklad 3. Ze dvou výběrů rozsahu $r = 228$ jsme dostali koeficienty korelace ${}_1r_{xy} = 0{,}56$ a ${}_2r_{xy} = 0{,}65$. Máme testovati, je-li tento rozdíl významný. Použijeme k tomu nejprve z'-transformace, ale vzhledem k značnému rozsahu můžeme také použíti k přibližnému řešení směrodatné odchylky (114).

Pro test z' dostáváme

$$z'_1 = \tfrac{1}{2}\{\lg(1+0{,}56) - \lg(1-0{,}56)\} = \tfrac{1}{2}\{\lg 1{,}56 - \lg 0{,}44\} =$$
$$= \tfrac{1}{2}\lg\frac{1{,}56}{0{,}44} = 0{,}6328,$$

$$z'_2 = \tfrac{1}{2}\{\lg(1+0{,}65) - \lg(1-0{,}65)\} = \tfrac{1}{2}\{\lg 1{,}65 - \lg 0{,}35\} =$$
$$= \tfrac{1}{2}\lg\frac{1{,}65}{0{,}35} = 0{,}7753,$$
$$z'_2 - z'_1 = 0{,}1425,$$
$$\sigma_{z_2'-z_1'} = \sqrt{\tfrac{1}{225} + \tfrac{1}{225}} = \tfrac{1}{15}\sqrt{2} = 0{,}0943.$$

Rozdíl je tedy menší než dvojnásobek směrodatné odchylky, takže jej nepovažujeme za významný. Na znaménko koeficientu korelace nebéřeme ve formulích pro z' zřetel, poněvadž testujeme numerický rozdíl mezi koeficienty korelace.

Úloha: Odvoďte podle rovnice (118)

a) výraz $r_{xy} = (e^{2z'} - 1) : (e^{2z'} + 1)$,

b) dokažte, že $r_{xy} = thz'$.

(9,4) Testování významnosti koeficientu regrese. — K provedení testu významnosti koeficientu regrese můžeme použíti t-testu. Tak jako v případech testování významnosti dřívějších charakteristik také zde si představujeme body znázorňující prvky se zjištěnou dvojicí znaků x, y a příslušné regresní přímky jako obraz dvojrozměrného roztřídění výběru z nějakého základního souboru — ať je to nějaký skutečný větší soubor nebo jen myšlenkově možný, který si sestrojíme k tomu cíli — abychom mohli významnost svými logickými pravidly přezkoušeti. V tomto základním souboru existuje také příslušná přímka regrese, jejímž přibližným odhadem je přímka určená z r dvojic pozorování na výběru tohoto rozsahu a jako taková má tedy svůj obor náhodných odchylek. Ptáme se proto, zda můžeme se svým stupněm statistické bezpečnosti souditi, že s rostoucími hodnotami proměnné x rostou (a v případě záporného koeficientu korelace klesají) průměrně hodnoty y. Abychom dali odpověď na tuto otázku, vyjdeme od t. zv. nulové hypotésy, že hodnota koeficientu regrese v základním souboru je nula, čili, že přímka regrese v základním souboru je rovnoběžná s vodorovnou osou x. To je pak totéž, jako kdybychom předpokládali, že proměnné v základním souboru ne-

jsou ve vztahu, čili $r(x, y) = 0$ a tím je test převeden na případ odstavce (9,1).

Lze však odvoditi nové vyjádření tohoto testu. K tomu cíli tedy stanovíme rozptyl koeficientu regrese proměnné y vzhledem k x, který je vyjádřen formulí (95), již napíšeme pomocí (77) ve tvaru

$$b_{21} = r_{xy}\frac{\sigma_y}{\sigma_x} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \cdot \frac{\sqrt{\Sigma(y-\bar{y})^2}}{\sqrt{\Sigma(x-\bar{x})^2}} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}.$$

Čitatele tohoto zlomku však můžeme ještě upraviti, neboť

$$\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - r\bar{x}\bar{y} = \Sigma xy - \bar{x}\Sigma y = \Sigma y(x-\bar{x})$$

a tedy

$$b_{21} = \frac{\Sigma y(x-\bar{x})}{\Sigma(x-\bar{x})^2}. \tag{120}$$

Abychom pak stanovili výběrový rozptyl této charakteristiky b, budeme uvažovati základní soubor, jehož prvky jsou náhodné výběry, které všechny mají tytéž hodnoty proměnné x. Stanovíme-li tedy z každého výběru charakteristiku b, budou odchylky jejích hodnot (nebo jinými slovy variace její) způsobeny jen tou okolností, že pro určitou hodnotu proměnné x nejsou hodnoty y v základním souboru, z něhož bereme výběry, všechny stejné. Označíme Y hodnoty proměnné y, které vyplývají z rovnice přímky regrese (92), jakožto přímky odhadu, čili píšeme rovnici přímky regrese

$$Y = \bar{y} + b_{21}(x-\bar{x}). \tag{121}$$

Průměrnou čtvercovou odchylku hodnot y od této přímky odhadu pro určitou danou hodnotu x pak označíme s a její čtverec, analogický rozptylu, tedy s^2. V čitateli (120) však je každá odchylka y od regresní formule základního souboru

násobena $(x - \bar{x})$, takže rozptyl tohoto součinu je $s^2 (x - \bar{x})^2$ a rozptyl celého součtu těchto součinů je $s^2 \Sigma (x - \bar{x})^2$. Poněvadž ve jmenovateli je $\Sigma (x - \bar{x})^2$, jehož rozptyl je $\{\Sigma (x - \bar{x})^2\}^2$, dostaneme pro celkový výběrový rozptyl charakteristiky b výraz

$$\frac{s^2 \Sigma (x - \bar{x})^2}{\{\Sigma (x - \bar{x})^2\}^2} = \frac{s^2}{\Sigma (x - \bar{x})^2}.$$

Průměrnou čtvercovou odchylku hodnot y od přímky regrese (121), t. j. od vypočítaných hodnot Y odhadneme, dělíme-li součet čtverců $(y - Y)^2$ počtem stupňů volnosti $n = r - 2$, neboť byly z výběrů rozsahu r již určeny dvě charakteristiky $\bar{y}$ a b_{21}, jež vstupují do výpočtu hodnot Y. Bude tedy

$$s = \sqrt{\frac{\Sigma (y - Y)^2}{r - 2}} \tag{122}$$

a rozptyl koeficientu regrese tudíž je

$$s_b = \frac{s}{\sqrt{\Sigma (x - x)^2}}. \tag{123}$$

Test významnosti můžeme provésti pomocí t-rozdělení, takže vypočítáme

$$t = \frac{b_{21}}{s_b} = \frac{b_{21} \sqrt{\Sigma (x - \bar{x})^2}}{s} \tag{124}$$

a tabulky 5 použijeme při $n = r - 2$ stupních volnosti. Způsob výpočtu je proveden v následujícím příkladě.

Oprávněnost tohoto postupu ukážeme objasněním, že je to totéž t, kterého užíváme při testování koeficientu korelace.

Vzhledem k (122) můžeme psáti

$$t^2 = b_{21}^2 \Sigma (x - \bar{x})^2 \frac{r - 2}{\Sigma (y - Y)^2}$$

a $\Sigma (y - Y)^2$ můžeme nahraditi podle (76) výrazem $(1 - r_{xy}^2) \Sigma (y - \bar{y})^2$, takže dostaneme

$$t^2 = b_{21}{}^2 \frac{\Sigma(x - \bar{x})^2}{\Sigma(y - \bar{y})^2} \frac{r - 2}{1 - r_{xy}{}^2} = r_{xy}{}^2 \frac{r - 2}{1 - r_{xy}{}^2}$$

vzhledem k rovnici (95). Vidíme tudíž, že

$$t = \frac{r_{xy}\sqrt{r - 2}}{\sqrt{1 - r_{xy}{}^2}}, \tag{125}$$

což je totéž jako v rovnici (117).

Výrazu pro rozptyl koeficientu regrese je možno užíti netoliko k testování nulové hypotésy, nýbrž také v případě, je-li koeficient regrese v základním souboru $b(y, x) \neq 0$ známý. Tuto hypotésu tedy testujeme tak, že použijeme t-testu, abychom ukázali zda je rozdíl $b_{21} - b(y, x)$ dělený svou směrodatnou odchylkou, významný při $r - 2$ stupních volnosti.

Máme-li testovati zda dva výběry, pro něž jsme dostali dva různé koeficienty regrese, jsou z jednoho základního souboru nebo ze základních souborů s týmž koeficientem regrese, založíme test významnosti rozdílu mezi dvěma koeficienty regrese na jejich směrodatných odchylkách. Jsou-li s'_b a s''_b počítány podle rovnice (123), je směrodatná odchylka rozdílu $b'_{21} - b''_{21}$ dána $s_{b'-b''} = \sqrt{s'^2_b + s''^2_b}$, takže

$$t = \frac{b'_{21} - b''_{21}}{s_{b'-b''}} \tag{126}$$

Tyto koeficienty regrese mohou býti počítány z výběru různého rozsahu $r_1 \neq r_2$, takže celkový počet stupňů volnosti je $n = n_1 + n_2 = r_1 - 2 + r_2 - 2$.

(9,4,1) Příklad. Jest testovati významnost koeficientu regrese v příkladu (7,2,1,1), str. 123.

Abychom mohli testovati pomocí rovnice (124), musíme počítati s, což se nejlépe počítá na základě identity

$$\Sigma(y - Y)^2 = \Sigma(y - \bar{y})^2 - b_{21}{}^2\Sigma(x - \bar{x})^2. \tag{127}$$

Tuto identitu nejprve dokážeme. Z rovnice (121) vyplývá

odečtením každé strany od y

$$y - Y = y - \bar{y} - b_{21}(x - \bar{x}),$$

takže umocněním na druhou dostaneme rovnici ve tvaru

$$(y - Y)^2 = (y - \bar{y})^2 + b_{21}^2 (x - \bar{x})^2 - 2b_{21}(y - \bar{y})(x - \bar{x}).$$

Provedeme-li nyní součet přes všechny hodnoty proměnných x a y, dostáváme

$$\Sigma(y - Y)^2 = \Sigma(y - \bar{y})^2 + \\ + b_{21}^2 \Sigma(x - \bar{x})^2 - 2b_{21}\Sigma(y - \bar{y})(x - \bar{x}) \qquad (128)$$

a poslední člen můžeme rozvésti

$$-2b_{21}\Sigma(y - \bar{y})(x - \bar{x}) = \\ = -2b_{21}\Sigma y(x - \bar{x}) + 2b_{21}\bar{y}\Sigma(x - \bar{x}).$$

Poněvadž podle (120) je $b_{21}\Sigma(x - \bar{x})^2 = \Sigma y(x - \bar{x})$, bude první člen pravé strany $-2b_{21} \cdot b_{21}\Sigma(x - \bar{x})^2$ a druhý člen se rovná nule, neboť obsahuje součet odchylek od průměru. Dosadíme-li tyto výsledky do (128), vidíme, že platnost identity (127) je prokázána.

Testujeme tedy významnost koeficientu regrese v regresní přímce $\eta = 5{,}77\xi$, takže stanovíme hodnotu

$$t = b_{21} \frac{\sqrt{\Sigma(x - \bar{x})^2}}{s}.$$

Nejprve je

$$\sqrt{\Sigma(x - \bar{x})^2} = \sqrt{52{,}81} = 7{,}267.$$

Poněvadž

$$s = \sqrt{\frac{1}{r - 2}[\Sigma(y - \bar{y})^2 - b_{21}^2\Sigma(x - \bar{x})^2]},$$

stanovíme

$$b_{21} = 5{,}77, \quad b_{21}^2 = 33{,}293,$$
$$\Sigma(y - \bar{y})^2 = 3553$$
$$b_{21}^2\Sigma(x - \bar{x})^2 = 1758$$
$$1795 : 154 = 11{,}65; \quad s = 3{,}413,$$

$$t = 5{,}77 \frac{7{,}267}{3{,}413} = 12{,}28,$$

a tato hodnota svědčí o vysoké významnosti, porovnáme-li ji s hodnotami tab. 5, ať na pěti- či jednoprocentní hranici významnosti.

(9,5) Test významnosti $r_{y.zz}$ a $r_{xy.z}$. Je třeba ještě, abychom se zmínili o testování významnosti koeficientu mnohonásobné korelace a koeficientu dílčí korelace.

K testování koeficientu dílčí korelace lze užíti t-testu týmž způsobem jako k testování jednoduchého koeficientu korelace, jen je třeba stanoviti správně počet stupňů volnosti. Není zde $n = r - 2$, nýbrž se zmenší o počet proměnných, který považujeme při výpočtu koeficientu dílčí korelace za konstantní; označíme-li jej k, bude pak $n = r - k - 2$ a tedy

$$t = \frac{r_{xy.z\ldots}}{\sqrt{1 - r_{xy.z\ldots}^2}} \sqrt{r - k - 2}. \qquad (129)$$

Pro testování významnosti koeficientu mnohonásobné korelace je třeba zvláštní tabulky, neboť je větší než každý z koeficientů jej vytvářejících a jeho minimum není -1, nýbrž 0. Postup obdobný těm, které jsme dosud vyložili, vede k chybným výsledkům; užívá se proto prostředků, jež poskytuje analysa rozptylu, o níž se budeme moci zmíniti v této knížce jen náznakem.

(10) Analysa rozptylu.

Uvažovali jsme dosud rozptyl určitého znaku jako charakteristiku nějakého souboru v celku. Přihlédneme-li k tomu, že podle některého znaku není soubor homogenní a skládá se třeba z jistých oblastních skupin, z nichž některé mají menší rozptyl a některé větší než je zjištěný celkový rozptyl, pak můžeme usuzovati skoro bezpečně, že i v jednotlivých skupinách není rozptyl přesně homogenní, leč že by byl ryze náhodný, což znamená způsobený velikým množstvím malých příčinných činitelů, z nichž jeden od druhého nelze ro-

zeznati. Je proto důležito při studiu rozptylu najíti možnost odlišovati rozptyl podle příčin nebo podle jejich skupin. Tuto možnost poskytuje analysa rozptylu, která podává výsledky, na něž můžeme užíti testů významnosti.

Poukázali jsme již v příkladě (3,7,1) na str. 54 na to, že v řadě pozorování, v níž jednotlivé hodnoty znaku vykazují jen náhodné odchylky, mohou průměry a rozptyly částečných souborů vzniklých nějakým roztříděním vykazovati také jen náhodné odchylky. Není-li však materiál stejnorodý, nýbrž vyskytují-li se při nějakém roztřídění podstatné rozdíly, budou směrodatné odchylky vypočítané z těchto částečných souborů větší, což můžeme statisticky zjistiti.

V nejjednodušší formě bývá užíváno tohoto postupu:

Stanovíme rozptyl celého uvažovaného souboru rozsahu r, při čemž použijeme podle rovnice

$$\sigma^2(x, v) = \frac{\sum_{j=1}^{r} \zeta_j}{r - 1}$$

(str. 41) $r - 1$ stupňů volnosti, takže bude

$$\sigma_x^2 = \frac{1}{r - 1} \Sigma(x - \bar{x})^2. \tag{130}$$

Sestává-li soubor z několika k částečných souborů, které mají rozsah resp. $\nu_1, \nu_2, \ldots, \nu_k$, pro něž máme zkoumati existenci podstatných rozdílů, stanovíme součty čtverců odchylek od průměrů $\bar{x}_i$ v každém z k částečných souborů a provedeme odhad směrodatné odchylky pomocí $r - k$ stupňů volnosti, neboť jsme z výběru vypočítali k průměrů. Tento odhad tedy bude

$$\sigma_{x,k}^2 = \frac{1}{r - k} \{\Sigma_1(x - \bar{x}_1)^2 + \Sigma_2(x - \bar{x}_2)^2 + \ldots + \Sigma_k(x - \bar{x}_k)^2\}, \tag{131}$$

kde se součet vztahuje vždy na všechny hodnoty proměnné v dotyčném částečném souboru.

Konečně pak můžeme provésti odhad rozptylu pomocí odchylek k průměrů částečných souborů od celkového průměru, takže je $k - 1$ stupňů volnosti. Jedná-li se jen o náhodné odchylky, bude tento odhad

$$\sigma_{\bar{x},k}^2 = \frac{1}{k-1}\{\nu_1(\bar{x}_1 - \bar{x})^2 + \nu_2(\bar{x}_2 - \bar{x})^2 + \ldots + \nu_k(\bar{x}_k - \bar{x})^2\}. \tag{132}$$

Tímto způsobem byl celkový rozptyl souboru rozložen na složku rozptylu „mezi skupinami" (132) a rozptyl „uvnitř skupin" (131).

Lze ukázati, že součet čtverců ve velkých závorkách rovnic (131) a (132) dává dohromady součet čtverců v (130).

Pro jednoduchost provedeme důkaz v případě, že částečné soubory, jimž také říkáme variety, mají stejný rozsah, t. j. $\nu_1 = \nu_2 = \ldots = \nu_k = \nu$ a tudíž $r = k\nu$. Odchylku pozorovaného znaku každého prvku od celkového průměru $x - \bar{x}$ můžeme rozložit ve dvě složky, v odchylku od průměru variety $x - \bar{x}_i$ a v odchylku tohoto průměru od celkového průměru $\bar{x}_i - \bar{x}$, jak je zřejmo z rovnice

$$x - \bar{x} = x - \bar{x}_i + \bar{x}_i - \bar{x}. \tag{133}$$

Utvoříme čtverce těchto odchylek

$$(x - \bar{x})^2 = (x - x_i)^2 + (\bar{x}_i - \bar{x})^2 + 2(x - \bar{x}_i)(x_i - x) \tag{134}$$

a sečteme pro všechny prvky jedné variety

$$\Sigma(x - \bar{x})^2 = \Sigma(x - \bar{x}_i)^2 + \nu(\bar{x}_i - \bar{x})^2; \tag{135}$$

člen $2(\bar{x}_i - x)\Sigma(x - \bar{x}_i) = 0$, neboť obsahuje jako součinitel součet odchylek hodnot znaku v jedné varietě od jejich průměru, který se tudíž rovná nule. Rovnic (135) máme k pro $i = 1, 2, \ldots, k$. Sečteme-li je všechny, dostaneme na levé straně součet čtverců odchylek v celém pozorovaném souboru, který bude

$$\Sigma\Sigma(x - \bar{x})^2 = \sum_i \Sigma(x - \bar{x}_i)^2 + \nu\sum_i(\bar{x}_i - \bar{x})^2 \tag{136}$$

a tu vidíme, že první součet na pravé straně zahrnuje členy velké závorky (131) a druhý zahrnuje členy velké závorky v (132), čímž je důkaz proveden. Kdyby se jednalo o veliké soubory, takže také rozsah variet v by byl velký, počítali bychom rozptyl tak, že součet čtverců odchylek od průměru bychom dělili jejich počtem. Dělíme-li tedy celou rovnici (136) součinem kv, dostaneme na levé straně rozptyl σ_x^2 a na pravé straně bude první člen $\frac{1}{k}\Sigma\sigma_{x,i}^2$ průměrným rozptylem všech k variet, označíme-li $\sigma_{x,i}^2$ rozptyl i-té variety a druhý člen $\frac{1}{k}\Sigma(\bar{x}_i - \bar{x})^2$ je rozptylem průměrů variet kolem celkového průměru.

Máme-li však co činiti s malými soubory, jak tomu většinou bývá v případech užívání analysy rozptylu, pak užíváme k výpočtu rozptylů příslušných stupňů volnosti, jak jsme učinili v rovnicích (130), (131), (132) a pro ně platí zřejmě rovnice

$$r - 1 = r - k + k - 1.$$

Je-li soubor homogenní, takže variety nemají rozptyly podstatně se od sebe lišící, nýbrž odchylky jsou jen náhodné, pak výrazy (131) a (132) jsou odhadem rozptylu celého souboru. Významnost odchylek lze pak opět testovati pomocí z testů (str. 66), s nimiž se čtenář může blíže seznámiti v [1].

Doslov.

Podali jsme stručně systém teoretické statistiky s vyloučením těch oborů, které se týkají časových řad a kinematiky vůbec. Již v tomto výkladu se ukázala mnohostrannost a složitost úvah, jež potřebuje moderní statistika, která je sice mladou vědou, ale již tak bohatě rozvinutou, že pole jejího používání je téměř nepřehledné. Současně je zřejmo, že nemůže býti nikdo jen se znalostí čtyř základních početních operací skutečným statistikem.