Jak vytváří statistika obrazy světa a života. II. díl

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

Persistent URL: http://dml.cz/dmlcz/403055

Terms of use:

© Jednota českých matematiků a fyziků

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://dml.cz

CESTA K VĚDĚNÍ SV. 26

Prof. Dr Jaroslav Janko

JAK VYTVÁŘÍ STATISTIKA OBRAZY SVĚTA A ŽIVOTA

II. DÍL

J č
m+f

Prof. Dr. Jar. Janko:

JAK VYTVÁŘÍ STATISTIKA OBRAZY SVĚTA A ŽIVOTA

II. díl.

V praktickém životě, v obchodě, v továrnách, je zvykem usuzovati na jakost zboží a výrobků na základě vzorků, neboli jak se ve statistice říká na základě náhodného výběru. Náhodný výběr je také vydatnou pomůckou při hledání zákonů v přírodních vědách. Jeho theorie je logickým základem pro odvozování soudů, poznatků a vět platných pro *celé* soubory, z nichž byl výběr vzat; toto odvozování se děje, jak už autor vyložil v prvním díle, *statistickou indukcí* z poznatků získaných z náhodných výběrů.

V tomto druhém svazku věnoval svou pozornost znaku kvantitativnímu. Nejdříve na *známém* základním souboru ukazuje na hlubší význam momentů rozdělení četnosti výběrových průměrů, a jiných výběrových charakteristik atd. Potom teprve se zabývá náhodnými výběry z *neznámého* základního souboru, osvětluje čtenáři význam zavedených charakteristik, odhaduje průměry,

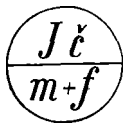
C E S T A K V Ě D Ě N Í

PROF. DR. JAROSLAV JANKO

JAK VYTVÁŘÍ STATISTIKA OBRAZY SVĚTA A ŽIVOTA

DÍL II.

S 18 obrazci v textu



Vyšlo jako 26. svazek sbírky

CESTA K VĚDĚNÍ

vydávané Jednotou českých matematiků a fysiků v Praze za redakce Dra R. BRDIČKY, Dra M. A. VALOUCHA a Dra F. VYČICHLA

1 9 4 4

NÁKLADEM JEDNOTY ČESKÝCH MATEMATIKŮ A FYSIKŮ
V GENERÁLNÍ KOMISI NAKLADATELSTVÍ PROMETHEUS V PRAZE
TISKEM KNIHTISKÁRNY „PROMETHEUS" V PRAZE VIII

Veškerá práva vyhrazena.

V prvním dílu, který vyšel jako 22. svazek této sbírky, jsme se seznámili s obecnou teorií statistiky, jež zahrnuje na prvním stupni vymezení statistické jednotky pro sestrojení určitého statistického souboru, jenž má býti podroben šetření podle jistých znaků. Na druhém stupni jsou soubory základní a soubory odvozené vyjádřeny čísly udávajícími jejich rozsah, čímž je dán podklad pro sestavení tabulek podle potřebných hledisek. Třetí stupeň tvoří početní zpracování, jehož cílem je stanovení charakteristik a hledání vztahů tvořením skupin a srovnáváním řad. Tyto metody jsou nezávislé na specielním oboru, v němž jsou aplikovány. K obecné teorii se připojuje potom specielní teorie, která se zabývá přizpůsobením obecné teorie pro užití v jednotlivých oborech (hospodářství, demografie, přírodních věd, . . .). Je tak obsáhlá, že by vyžadovala zvláštních svazků, neboť musejí býti rozšířeny na př. metody obecné teorie statistiky pro aplikaci na hospodářský život o nauku o indexních číslech, o metody výpočtu trendu a různých cyklických variací, především sezonních; pro aplikaci v demografii je třeba rozšíření o nauku o konstrukci tabulek úmrtnosti, o výpočtu čísel reprodukčních, o metodách vyrovnávacích, interpolaci a pod.

Statistika se zabývá soubory viditelných a takřka hmatatelných předmětů, na nichž zjišťuje potřebné znaky; vychází tedy ze zkušenosti empirickému názoru přístupné, kterou svým způsobem popíše. Vyšším stupněm statistické práce pak je metodika odhadu. Ve většině případů se stává statistika pro poznávání výseků života a přírody užitečnou teprve metodou odhadu a to zvláště tam, kde je třeba vyhnouti se přílišným nákladům na statistické šetření, obtěžování obyvatelstva nebo kde je vyčerpávající šetření znemožněno nepřístupností celého souboru. Tak vznikají úkoly jak odhadovati vlastnosti velkého souboru, který z pozorování

neznáme, podle souboru menšího rozsahu, který jej zastupuje čili representuje. Usuzování z části na celek je umění provozované odedávna. Společnost pak se vždy ptala po předpovědi a žádala ji od statistika, kterého při dřívějších málo vyvinutých metodách „přepadal nepříjemný pocit, měl-li vstoupiti na kolísavou palubu odhadů a ještě často dnes se mu zatají dech, je-li unesen do výšin kombinací a dedukcí, které mu dávají místo čísel do posledního místa přesně zjištěných jen řádové odhady nebo pravděpodobnosti".

Věda dneška měří a předpovídá; jsou v ní také rozšířeny předpovědi, jejichž splnění je podmíněno uskutečněním předpokladů, které nejsou přesně popsány. Je třeba metod, jimiž by se mohla zkoušeti jejich správnost a měřit jejich pravděpodobnost. Maxwell proslovil skvělé paradoxon, že teorie pravděpodobnosti je jedinou logikou praktického muže. Ovšem průměrný občan nemá dostatečný výcvik v počítání nebo odhadování pravděpodobností. Počet pravděpodobnosti umožňuje zkoušeti předpovědi na jejich náhodnou povahu, jak ve výkladech tohoto svazku uvidíme. Objeví se také, že nejbezpečnější předpovědi spočívají na úplné náhodnosti, neboť pravděpodobnost jejich uskutečnění lze přesně určiti. Seznáme totiž v náhodnosti zvláštní formu pravidelnosti, takže v ní není nic tajemného, ježto má své zákonitosti. Formule k jejím výpočtům potřebné vznikly ovšem za určitých předpokladů a užívá se jich tedy s jistými výhradami. Statistik užívá v konkretním případě jen těch, kterých je právě k sledovanému účelu nutně třeba. Musí dobře ovládat pracovní metody matematiky, která je tu jen pomocnou vědou, aby se nestal otrokem početního mechanismu. Především musí statistik provádějící větší statistickou práci znáti prakticky obor, v němž zkoumání provádí, aby měl bezprostřední poměr k původnímu materiálu. Formulí pak smí užívati jen tehdy, rozumí-li tomu jak vznikly, jejich správnému významu a předpokladům jejich platnosti.

Ve statistice má místo také dekorativní umění, kterého jsme užili i v tomto druhém dílu statistiky. Je to grafické znázornění, které slouží většinou popularisování statistiky. Víme, že deset čísel můžeme snadno obsáhnouti, ale dvacet jen s námahou a sto již vůbec ne. Mnohému čtenáři jsou tudíž tabulky a texty promíšené číslicemi neztravitelné, kdežto grafikon mluví k srdci; bývá trochu povrchní, ale vlichotí se. Zodpovědnost vůči čtenáři vyžaduje, aby se nevyskytovaly formule jako dekorace pojednání bez vztahu k účelu prováděného zkoumání. Chceme proto na tomto místě, třeba malého rozsahu, umožniti proniknutí k základům statistické indukce odvozením hlavních formulí teorie náhodného výběru, jichž užívá representativní metoda a tedy osvětlením jejich vzniku. Jen tam uvádíme pouze výsledky, kde vznikají zcela analogicky jako předcházející nebo kde by odvození přesahovalo rámec sledovaný touto sbírkou.

Také v tomto druhém dílu, který podává úvod do matematiky representativní metody jsme se museli omeziti jen na metody nejdůležitější, z nichž velkou část zaujal výklad o korelaci. Již v prvním dílu jsme vyložili podstatu teorie náhodného výběru a její hlavní problémy; řešení však bylo provedeno jen pro znak alternativní. Tento druhý díl tedy podává výklad teorie náhodného výběru pro znak kvantitativní a seznamuje čtenáře s pojmy zavedenými do statistiky v poslední době velikého rozvoje. Teorie korelace se tu podává od přístupného výkladu základních pojmů až do praktického užití a hodnocení charakteristik vzhledem k náhodným odchylkám výběrovým. Výklady tohoto druhého dílu jsou jen prvními úvodními kroky do teorie statistiky právě tak, jako tomu bylo v prvním dílu, třebaže v některých místech musejí klásti větší požadavky na pozornost čtenářovu. Tyto kroky pečující o tvoření čistých pojmů mohou otevřít pohled na mnohostrannost problémů a chrániti před nebezpečím, které vzniká používáním nevhodného nástroje při práci. Aplikaci těchto metod statistiky se ote-

vírá takřka nekonečné pole bádání, ale v každém zvláštním oboru je pak třeba vypracovati ještě četné specielní metody. Největšího uspokojení dosáhne ten čtenář, který ovládaje tuto teorii může prožít aktivní spojení její s konkretním předmětem, t. j. s určitým pozorovaným souborem prvků.

Na konec děkuji opět p. doc. dr. Fr. Vyčichlovi, redaktoru této sbírky za laskavé opatření obrázků a Jednotě českých matematiků a fysiků za úpravné vydání knížky.

V Praze v červnu 1942.

Jaroslav Janko.

ČÁST I.

TEORIE NÁHODNÝCH VÝBĚRŮ.
(Znak kvantitativní.)

(I) Úvod.

V každodenních záležitostech všedního života a v pracech svého zaměstnání jsme vedeni k tvoření úsudků, které zevšeobecňují poznatky získané na jistém omezeném počtu pozorovaných případů. Když hospodyně koupila deset housek v určitém obchodě a shledala, že pět z nich není tak čerstvých jak by si bylo přáti, rozhodne se, že je bude příště kupovati jinde. Když někdo čekal pětkrát v týdnu deset minut nebo déle na vůz elektrické dráhy na stanici, činí závěr, že dopravní možnost je poměrně chudá. Lékař vezme kapku krve pacientovy, rozředí ji a počítá pak pod mikroskopem krvinky v nepatrné částce zředěného roztoku, aby na podkladě tohoto materiálu poznal podstatné vlastnosti krve pacientovy pro svou diagnosu. Prodává-li se pšenice podle toho, jak je bohatá na lepek, posuzuje se podle malých množství jako vzorků vzatých z celé sklizně. Také obchod jiným zbožím se provádí odedávna pomocí vzorků. Jsou tedy v praktickém životě časté případy, kdy je nutno usuzovati na vlastnosti jistého souboru prvků, jež nemůžeme všechny pozorovat, na základě toho, co jsme zjistili na nějakém menším množství prvků z něho, nebo jak často říkáme na vzorku. Dospěli jsme zkušeností a intuicí k této víře, že nám může vzorek něco povědětí o celém základním množství, z něhož byl vzat a že nám to může povědětí tím lépe, čím je větší. Jestliže výrobce ložiskových kuliček kontroluje své výrobky na nejvyšší možnou zatížitelnost podle vzorků a usuzuje z toho na jakost celé výroby, činí totéž co experimentátor, který vychází z předpokladu, že je možno usuzovati z následků na příčiny a z pozorovaných zvláštních případů odvozovati věty obecněji platné. Říkáme

tomu v logice indukce čili závěr postupující od zvláštního k obecnému nebo úsudek z výběru na základní soubor.

Je to úsudek, kterým se rozšiřují výsledky odvozené z pozorovaného souboru na rozsáhlejší soubor obsahující prvky, jež nebyly v původně studovaném souboru. Soustředíme-li informaci o pozorovaném souboru do několika charakteristik, můžeme pak usuzovati podle těchto hodnot na hodnoty příslušných parametrů v rozsáhlejším základním souboru. Metody, jimiž docilujeme zobecnění statistických výsledků odvozených z výběrů, nazýváme statistickou indukcí.

Bylo by možno uvésti velkou řadu případů, kdy je možno nebo nutno dosáhnouti veškeré informace o základním souboru podle jednoho malého náhodného výběru či několika málo výběrů. Víme, že se mnohdy provede ve fysice osm až dvanáct pokusů, aby se odvodil přírodní zákon určitého jevu; nebo botanik zkoumá charakteristiky malého výběru několika rostlin určitého druhu a přisuzuje pak tomuto druhu vlastnosti, které získal z výběru. Podobně závisí předpovídání vývoje obchodu na informaci, která vyplyne z výběru. Možno říci, že skoro každá empirická formule je získána pomocí výběrových dat. Důvody užívání těchto metod mohou spočívati také v nutnosti úspory času a nákladu, jehož vyžadují taková šetření a zkoušky. Při zkoušení nebo kontrole některých vlastností předmětů musí býti dotyčné předměty zničeny nebo znehodnoceny jako je tomu na př. při zkoušení citlivosti fotografických desek, délky života žárovek, pevnosti trubek a pod.

Připouštíme zajisté, že tento postup od zvláštního k obecnému musí obsahovati jakési prvky nejistoty. To ovšem neznamená, že závěry statistické indukce nejsou zcela přesné, ježto máme prostředky, které jsou s to vyjádřiti přesně povahu a stupeň oné nejistoty. Doklady toho máme v aplikaci teorie pravděpodobnosti. Jednotlivý případ je sice nejistý, ale mohou býti odvozeny přesnou dedukcí pravděpodobnosti různých možných jevů nebo jejich kombinací.

Skutečnost, že v úsudcích, k nimž jsme dospěli indukcí, je jakási nejistota, nevylučuje možnost zcela přesných a jednoznačných závěrů. V interpretaci odvozených pravděpodobností mívá někdy původ nejistota, s níž se setkáváme. Logický základ a soustavu měr pro vyvozování úsudků statistickou indukcí nám poskytuje teorie náhodného výběru. Hlavní výsledky a aplikaci její v případě alternativního znaku jsme vyložili již v I. díle.

(2) Náhodné výběry ze známého základního souboru.

Abychom odvodili výsledky upotřebitelné co nejlépe v praxi, tedy za podmínek co nejméně omezujících, budeme nejprve zkoumati, jaké hodnoty charakteristik poskytují všechny možné výběry ze známého základního souboru a vyslovíme tedy konkretněji první úkol teorie výběru. Známe základní soubor, čili jeho rozdělení četností, které je vyjádřeno několika parametry. Budeme uvažovati první tři parametry (průměr, směrodatnou odchylku a šikmost). Tímto základním souborem je podána informace o nějakém studovaném jevu. Chceme nyní popsati tento jev charakteristikami všech výběrů určitého rozsahu, které mohou vzniknouti ze základního souboru. V našem případě to tedy budou výběrové průměry, směrodatné odchylky a šikmosti, jejichž rozdělení četností budeme zkoumati, abychom našli souvislost mezi nimi a momenty základního souboru.

Známe rozdělení četností uvažovaného základního souboru (první dva sloupce tab. 1).

Parametry tohoto rozdělení četností, odvozené z prvních tří momentů budou

$$\bar{x} = \mu'(x,1) = 7{,}0000, \quad \sigma(x) = \sqrt{\mu(x,2)} = 2{,}002,$$

$$\alpha(x,3) = \frac{\mu(x,3)}{\sigma^3(x)} = 0{,}0000. \qquad (1)$$

Tabulka 1.

x_i	$n(x_i)$	$\bar{x}_i$	$n(\bar{x}_i)$	$f(\bar{x}_i)$
(1)	(2)	(3)	(4)	(5)
0	1	6.625—6,674	0	0,000
1	2	6,675—6,724	2	0,005
2	9	6,725—6,774	12	0,030
3	28	6,775—6,824	20	0,050
4	66	6,825—6,874	22	0,055
5	121	6,875—6,924	46	0,115
6	175	6,925—6,974	57	$0,142_5$
7	197	6,975—7,024	66	0,165
8	175	7,025—7,074	55	$0,137_5$
9	121	7,075—7,124	50	0,125
10	66	7,125—7,174	40	0,100
11	28	7,175—7,224	15	$0,037_5$
12	9	7,225—7,274	11	$0,027_5$
13	2	7,275—7,324	3	$0,007_5$
14	1	7,325—7,374	1	$0,002_5$
Součet	1001	Součet	400	1,0000

Ze základního souboru tohoto rozdělení četností vezmeme 400 výběrů, z nichž každý bude rozsahu $r = 200$ prvků. Můžeme to provésti na př. tak, že napíšeme pro každý prvek základního souboru lístek (celkem 1001), na němž je poznamenána hodnota znaku a vytáhneme z nich 200 lístků, které tvoří jeden výběr. Můžeme postupovati buď tak, že každý jednotlivý lístek hned zase vrátíme zpět, nebo vrátíme teprve vytažených 200 najednou. Tuto okolnost však zatím necháme stranou.

Vypočítáme pro každý výběr rozsahu 200 prvků průměr výběrový $\bar{x}_i$, takže dostaneme 400 hodnot $\bar{x}_i\,(i = 1, 2, \ldots$ $\ldots, 400)$, které budeme považovati za prvky nového souboru rozsahu $n = 400$; je to tedy soubor výběrových průměrů. Každý z těchto výběrových průměrů můžeme považovati za odhad parametru $\bar{x}$ základního souboru. Soubor výběrových průměrů má své určité rozdělení četností, které

si uvedeme v třídních intervalech délky $h = 0{,}05$, jejichž střed označíme ${}_0\bar{x}_i$, takže dostaneme sloupec 3 a 4, tab. 1.

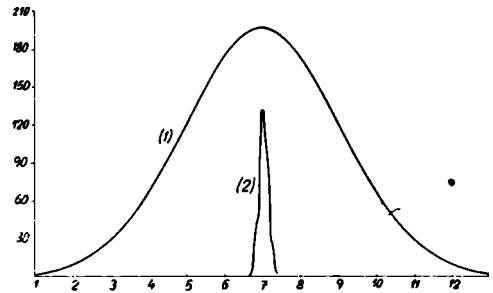
První tři charakteristiky tohoto rozdělení četností jsou

$$\bar{x}_P = 7{,}005, \quad \sigma_P = 0{,}121, \quad \bar{\alpha}_3 = 0{,}0001. \tag{2}$$

Porovnáme-li nyní výsledky, jež jsme dostali pro parametry základního souboru a pro charakteristiky souboru výběrových průměrů, vidíme, že průměr základního souboru 7,0000 je velmi blízko průměru výběrových průměrů 7,005. Směrodatná odchylka v základním souboru je 2,002, kdežto v rozdělení výběrových průměrů je jen 0,121. Tato malá hodnota směrodatné odchylky rozdělení výběrových průměrů ukazuje, že všechny výběrové průměry leží velmi blízko svého průměru a také průměru základního souboru. Variační obor proměnné x je v základním souboru $x_{15} - x_1 = 14$, kdežto pro rozdělení výběrových průměrů je pouze ${}_0\bar{x}_{15} - {}_0\bar{x}_1 = 0{,}70$, což opět ukazuje, jak blízko leží hodnoty proměnné v rozdělení výběrových průměrů kolem průměru jejich a kolem průměru základního souboru. V tomto příkladě jsou všechny výběrové průměry ve vzdálenosti nejvýše 0,35 od průměru základního souboru. Porovnáme-li tuto hodnotu se směrodatnou odchylkou $\sigma_P = 0{,}121$, vidíme, že podle počtu pravděpodobnosti ([1], str. 38) je téměř nemožno dostati průměr výběru rozsahu $r = 200$, aby se lišil od průměru základního souboru o 0,5. Ukazuje tudíž tento příklad, že průměr z dosti rozsáhlého výběru je asi tak dobrý jako průměr základního souboru.

Není-li tedy základní soubor znám, je průměr z jednoho výběru nebo z několika málo výběrů asi tak dobrý, jak je vůbec možno získat. Nedáme se proto do velké práce s obstaráváním dalších výběrů nebo rozsáhlejšího výběru. Výsledky a právě provedené úvahy vidíme snadno v grafickém znázornění relativních četností obou rozdělení (obr. 1). Kdybychom utvořili průměr všech možných průměrů výběrových, rovnal by se přesně průměru základního souboru, jak níže dokážeme.

Zcela obdobně bychom nyní mohli na tomto příkladě zjistiti charakteristiky souboru 400 výběrových rozptylů nebo směrodatných odchylek a souboru výběrových šikmostí. Je to však postup zcela obdobný, proto se budeme



Obr. 1. Rozdělení četností základního souboru (1). Rozdělení četností výběrových průměrů (2).

věnovati hned řešení obecnému, kde jej provedeme pro všechny tyto soubory výběrových charakteristik. Na předcházejícím příkladu vidíme tedy, že při numerickém popisu daného rozdělení četností musíme rozeznávat dvě hlediska. Jednak můžeme považovat popis daného rozdělení za cíl pro sebe, jednak jej můžeme uvažovati jako výběr, representující větší soubor, t. zv. základní. Obyčejně je důležité toto druhé hledisko, neboť v případech, kdy je buď nepraktické nebo nemožné pozorovati nebo měřiti studovaný znak na všech prvcích základního souboru, musíme z něho vzíti výběr a podle něho usuzovati na celý základní soubor.

Přistoupíme tedy k obecnému řešení otázky jak dobře výběr popisuje základní soubor za předpokladu, že základní soubor má rozsah N a jsou z něho brány výběry rozsahu r, při čemž tedy $r \leqq N$, takže můžeme dostati celkem $\binom{N}{r}$ různých výběrů. Každý z těchto výběrů má nějaký první mo-

ment čili průměr, takže soubor těch $\binom{N}{r}$ výběrů dává rozdělení četností všech výběrových průměrů. Podobně má každý výběr druhý moment, takže soubor druhých výběrových momentů má také své rozdělení četností; stejně je tomu pro třetí a další momenty výběrové. Bude tudíž naším úkolem nyní srovnati momenty počítané z výběru s momenty základního souboru.

Je-li N konečné, říkáme, že tvoříme výběry z konečného základního souboru (na př. obyvatelé nějakého státu). Je-li N nekonečné, tvoříme výběry z nekonečného základního souboru (barometrický tlak v různých bodech atmosféry). V mnohých případech je N tak veliké, že můžeme prakticky považovati základní soubor za nekonečný a do výsledků tím nepronikne chyba, která by musela býti uvažována. Jsou také případy, kdy nevíme bezpečně je-li studovaný základní soubor konečný či nekonečný, jak je tomu na př. u souboru hvězd. Statistikové se dříve zabývali hlavně případem nekonečného základního souboru, ježto algebraické výpočty zvláště pro rozdělení vyšších momentů z konečného základního souboru jsou značně složité. Poněvadž se zde omezujeme na rozdělení četností několika prvních momentů, odvodíme výsledky pro N konečné a necháme-li pak v nich růsti N do nekonečna, vyplynou snadno zjednodušené výsledky pro výběry z nekonečného základního souboru.

(2, I) Momenty rozdělení četností výběrových průměrů.

Průměr výběrových průměrů. Pro konečné N a r můžeme vybrati ze základního souboru, jehož prvky mají pozorované hodnoty proměnné $x_1, x_2, \ldots, x_N$, celkem $\binom{N}{r} = \nu$ různých výběrů, z nichž každý má r prvků ($r \leq N$) a hodnota x_i se v něm vyskytuje jen jednou. Můžeme je nějakým způsobem seřadit, takže dostaneme ν výběrových průměrů

$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i, \ldots, \bar{x}_\nu$. Bude tedy $\bar{x}_i = \frac{1}{r}\sum^{r,i} x_l$, kde součet $\sum^{r,i}$ značí součet všech r hodnot znaku i-tého výběru. Abychom stanovili jejich průměr $\mu'_1 = \frac{1}{\nu}(\bar{x}_1 + \bar{x}_2 + \ldots \bar{x}_\nu)$ uvážíme, že počet prvků ve všech ν výběrech je dohromady $\nu \,.\, r$, kdežto počet různých prvků je dán počtem prvků N základního souboru. Vystupuje tedy jeden určitý prvek v tomto celku všech výběrů $\frac{\nu r}{N}$ krát. To platí pro každý z N prvků základního souboru, takže součet všech hodnot znaku ve všech výběrech bude $(x_1 + x_2 + \ldots + x_N)\frac{\nu r}{N}$, a poněvadž je těchto prvků celkem νr, je tudíž průměr

$$\mu'_1 = (x_1 + x_2 + \ldots + x_N)\frac{1}{N} = \bar{x}. \tag{3}$$

Dostáváme tak první důležitou větu:

Průměr μ'_1, čili první moment výběrových průměrů $\bar{x}_i$ se rovná průměru základního souboru $\bar{x}$.

Rozptyl výběrových průměrů. Další charakteristikou rozdělení výběrových průměrů je jejich rozptyl μ_2. Druhý moment výběrových průměrů, vzhledem k jejich průměru $\bar{x}$ je dán podle definice výrazem

$$\mu_2 = \frac{1}{\nu}\sum_1^{\nu}(\bar{x}_i - \bar{x})^2 \tag{4}$$

kde $\bar{x}_i = \frac{1}{r}\sum^{r,i} x_j$, takže označíme-li odchylky hodnot znaku od průměru v základním souboru

$$x_j - \bar{x} = \xi_j \tag{5}$$

bude

$$\bar{x}_i - \bar{x} = \frac{1}{r}\left(\sum^{r,i} x_j - r\bar{x}\right) = \frac{1}{r}\sum^{r,i}\xi_j \tag{6}$$

a čtverec

$$(\bar{x}_i - \bar{x})^2 = \frac{1}{r^2}\Big(\sum^{r,i} \xi_j\Big)^2 = \frac{1}{r^2}\Big[\sum^{r,i} \xi_j^2 + 2\sum_{j,k}^{r,i} \xi_j \xi_k\Big], \quad j \neq k.$$

Součet $\sum_{j,k}^{r,i}$ značí součet všech členů utvořených tak, že bereme součiny všech proměnných ξ v i-tém výběru po dvou.

Dosadíme-li do (4), dostaneme

$$\mu_2 = \frac{1}{\nu} \cdot \frac{1}{r^2}\left[\frac{r}{N}\nu \sum_{j=1}^{N} \xi_j^2 + 2\frac{\binom{r}{2}\nu}{\binom{N}{2}} \sum_{j,k=1}^{N} \xi_j \xi_k\right], \quad j \neq k \qquad (7)$$

neboť sčítáme čtverce odchylek všech hodnot znaku ve všech ν výběrech, takže jako v předchozím případě průměru také zde každý z N čtverců hodnot odchylek znaku od průměru se bude vyskytovat $\frac{r\nu}{N}$ krát. Podobně uvažujeme v případě sčítání druhého členu hranaté závorky. Součet $\sum_{j,k}^{r,i}$ obsahuje $\binom{r}{2}$ členů, takže v součtu pro všech ν výběrů je $\binom{r}{2}\nu$ sčítanců. Celkem je $\binom{N}{2}$ různých podvojných součinů N hodnot znaku, takže každý určitý podvojný součin $\xi_j \xi_k$ se bude vyskytovati $\binom{r}{2}\nu : \binom{N}{2}$ krát.

Pro součet odchylek od průměru a jejich čtverců můžeme psáti rovnice

$$\sum_{j=1}^{N} \xi_i = 0, \quad \sum_{j=1}^{N} \xi_j^2 = N\,\mu(x, 2); \qquad (8)$$

čtverec první z těchto rovnic bude

$$\Big(\sum_{j=1}^{N} \xi_j\Big)^2 = \sum_{j=1}^{N} \xi_j^2 + 2\sum_{j,k=1}^{N} \xi_j \xi_k = 0 \quad (j \neq k),$$

takže

$$2\sum_{j,k=1}^{N}\xi_j\,\xi_k = -N\,\mu(x,2) \tag{9}$$

Budeme-li dále označovati

$$\vartheta_i = \frac{\binom{r}{i}}{\binom{N}{i}} = \frac{r\,(r-1)\,(r-2)\ldots(r-i+1)}{N\,(N-1)\,N-2)\ldots(N-i+1)}, \tag{10}$$

dostaneme rovnici pro druhý moment vzhledem k průměru ve tvaru

$$\mu_2 = \frac{1}{r^2}\,N\,.\,\mu(x,2)\,[\vartheta_1 - \vartheta_2] \tag{11}$$

nebo

$$\mu_2 = \frac{1}{r}\,\mu(x,2)\left(1-\frac{r-1}{N-1}\right). \tag{12}$$

Pro směrodatnou odchylku výběrových průměrů σ_P pak najdeme odmocněním

$$\sigma_P = \sigma(x)\sqrt{\frac{N-r}{r(N-1)}} = \sigma(x)\,\frac{1}{\sqrt{r\,\dfrac{N-1}{N-r}}}. \tag{13}$$

Z toho výsledku je patrna druhá věta:

Rozptyl a tudíž také směrodatná odchylka výběrových průměrů je menší než rozptyl či směrodatná odchylka základního souboru.

Je-li N velké, takže můžeme psáti N místo $N-1$, přejde výraz (13) ve tvar

$$\sigma_P = \sigma(x)\sqrt{\frac{N-r}{rN}} = \sigma(x)\sqrt{\frac{1}{r}-\frac{1}{N}}.$$

Šikmost. Abychom našli šikmost rozdělení četnosti výběrových průměrů, odvodíme nyní jejich třetí moment

vzhledem k průměru podobně jako v předcházejícím odstavci. Podle definice

$$\mu_3 = \frac{1}{\nu} \sum_1^{\nu} (\bar{x}_i - \bar{x})^3,$$

což vzhledem ku (5) a (6) dává pomocí multinomického teorému

$$\mu_3 = \frac{1}{\nu} \sum_1^{\nu} \frac{1}{r^3} [\sum^{r,i} \xi_j^3 + 3 \sum^{r,i} \xi_j^2 \xi_k + 6 \sum^{r,i} \xi_j \xi_k \xi_l]. \quad (14)$$

Multinomický teorém je zobecněním binomického teorému, podle něhož je

$$(\xi_1 + \xi_2)^n = \sum \frac{n!}{a!\, b!} \xi_1^a \xi_2^b,$$

kde $n = a + b$ a součet se vztahuje na všechna možná rozložení čísla n ve dva sčítance, z nichž každý je číslo celé nezáporné. Rozšíří-li se počet členů v závorce, lze dokázati, že platí

$$(\xi_1 + \xi_2 + \ldots + \xi_N)^n = \sum \frac{n!}{a!\, b!\, c! \ldots} \xi_1^a \xi_2^b \xi_3^c \ldots$$

Číslo n se nyní rozloží zase všemi možnými způsoby na N celých nezáporných sčítanců $n = a + b + c + \ldots$ a součet se vztahuje na všechna tato rozložení čísla n.

Zcela obdobnými úvahami jako při odvozování rovnice (7) provedeme součty, takže dostaneme

$$\mu_3 = \frac{1}{r^3} [\vartheta_1 \sum_{j=1}^{N} \xi_j^3 + 3\vartheta_2 \sum_{j,k=1}^{N} \xi_j^2 \xi_k + 6\vartheta_3 \sum_{j,k,l=1}^{N} \xi_j \xi_k \xi_l] \quad (15)$$

a tyto součty můžeme vyjádřit pomocí momentů základního souboru na základě vztahů plynoucích ze sčítání třetího řádu

$$\sum_{j=1}^{N} \xi_j^3 = N\mu(x, 3), \qquad \sum_{j=1}^{N} \xi_j^2 \sum_{k=1}^{N} \xi_k = 0 = \sum_{j=1}^{N} \xi_j^3 + \sum_{j,k}^{N} \xi_j^2 \xi_k,$$

$$\sum_{j,k=1}^{N} \xi_j \xi_k \sum_{l=1}^{N} \xi_l = 0 = \sum_{j,k=1}^{N} \xi_j^2 \xi_k + 3 \sum_{j,k,l=1}^{N} \xi_j \xi_k \xi_l,$$

čili

$$\sum_{j,k=1}^{N} \xi_j^2 \xi_k = -N\mu(x, 3), \quad 3 \sum_{j,k,l=1}^{N} \xi_j \xi_k \xi_l = N\mu(x, 3),$$

takže po dosazení do (15)

$$\mu_3 = \frac{1}{r^3} N\mu(x, 3) [\vartheta_1 - 3\vartheta_2 + 2\vartheta_3] \tag{16}$$

a rozvedeme-li výrazy v závorce podle (10), můžeme také po jednoduché úpravě psáti

$$\mu_3 = \frac{1}{r^2} \mu(x, 3) \frac{(N-r)(N-2r)}{(N-1)(N-2)}. \tag{17}$$

Pro šikmost

$$\bar{\varrho} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{\mu_2 \sigma_P},$$

tudíž plyne po dosazení z (17), (12), (13) výraz

$$\bar{\varrho} = \varrho(x) \frac{N-2r}{N-2} \sqrt{\frac{N-1}{r(N-r)}}. \tag{18}$$

Šikmost výběrových průměrů je menší než šikmost základního souboru.

Exces. Stejným postupem vypočítáme čtvrtý moment vzhledem k průměru $\bar{x}$, abychom mohli vyjádřiti exces rozdělení četností výběrových průměrů. Podle multinomického teorému bude

$$\mu_4 = \frac{1}{\nu} \sum_{i=1}^{\nu} \frac{1}{r^4} [\sum^{r,i} \xi_j^4 + 4 \sum^{r,i} \xi_j^3 \xi_k + 6 \sum^{r,i} \xi_j^2 \xi_k^2 +$$

$$+ 12 \sum^{r,i} \xi_j^2 \xi_k \xi_l + 24 \sum^{r,i} \xi_j \xi_k \xi_l \xi_m]$$

čili

$$\mu_4 = \frac{1}{r^4}\Big[\vartheta_1 \sum_{j=1}^{N} \xi_j^4 + 4\vartheta_2 \sum_{j,k=1}^{N} \xi_j^3 \xi_k + 6\vartheta_2 \sum_{j,k=1}^{N} \xi_j^2 \xi_k^2 + \\ + 12\vartheta_3 \sum_{j,k,l=1}^{N} \xi_j^2 \xi_k \xi_l + 24\vartheta_4 \sum_{j,k,l,m=1}^{N} \xi_j \xi_k \xi_l \xi_m\Big]. \qquad (19)$$

Použijeme nyní vztahů

$$\sum_{j=1}^{N} \xi_j^4 = N\mu(x, 4), \qquad \sum_{j=1}^{N} \xi_j^3 \sum_{k=1}^{N} \xi_k = 0 = \sum_{j=1}^{N} \xi_j^4 + \sum_{j,k=1}^{N} \xi_j^3 \xi_k,$$

$$\Big(\sum_{j=1}^{N} \xi_j^2\Big)^2 = \sum_{j=1}^{N} \xi_j^4 + 2 \sum_{j,k=1}^{N} \xi_j^2 \xi_k^2,$$

$$\sum_{j=1}^{N} \xi_j^2 \Big(\sum_{k=1}^{N} \xi_k\Big)^2 = 0 = \sum_{j=1}^{N} \xi_j^2 \Big(\sum_{k=1}^{N} \xi_k^2 + 2 \sum_{k,l=1}^{N} \xi_k \xi_l\Big) = \\ = \Big(\sum_{j=1}^{N} \xi_j^2\Big)^2 + 2 \sum_{j,k=1}^{N} \xi_j^3 \xi_k + 2 \sum_{j,k,l=1}^{N} \xi_j^2 \xi_k \xi_l,$$

$$\Big(\sum_{j=1}^{N} \xi_j\Big)^4 = 0 = \sum_{j=1}^{N} \xi_j^4 + 4 \sum_{j,k=1}^{N} \xi_j^3 \xi_k + 6 \sum_{j,k,=1}^{N} \xi_j^2 \xi_k^2 + \\ + 12 \sum_{j,k,l=1}^{N} \xi_j^2 \xi_k \xi_l + 24 \sum_{j,k,l,m=1}^{N} \xi_j \xi_k \xi_l \xi_m,$$

z nichž plyne

$$\sum_{j,k=1}^{N} \xi_j^3 \xi_k = -N\mu(x, 4), \quad 2 \sum_{j,k=1}^{N} \xi_j^2 \xi_k^2 = N^2\mu^2(x, 2) - N\mu(x, 4),$$

$$2 \sum_{j,k,l=1}^{N} \xi_j^2 \xi_k \xi_l = 2N\mu(x, 4) - N^2\mu^2(x, 2),$$

$$24 \sum_{j,k,l,m=1}^{N} \xi_j \xi_k \xi_l \xi_m = -6N\mu(x, 4) + 3N^2\mu^2(x, 2),$$

takže dosazením do (19) dostáváme

$$\mu_4 = \frac{1}{r^4} N\mu(x, 4)\,[\vartheta_1 - 7\vartheta_2 + 12\vartheta_3 - 6\vartheta_4] + \\ + 3N^2\mu^2(x, 2)\,[\vartheta_2 - 2\vartheta_3 + \vartheta_4]. \qquad (20)$$

Pro pátý moment výběrových průměrů vzhledem k jejich průměru bychom dostali

$$\mu_5 = \frac{1}{r^5} N\mu(x, 5)\,[\vartheta_1 - 15\vartheta_2 + 50\vartheta_3 - 60\vartheta_4 + 24\vartheta_5] + \\ + 10N^2\mu(x, 2)\,\mu(x, 3)\,[\vartheta_2 - 4\vartheta_3 + 5\vartheta_4 - 2\vartheta_5]. \qquad (21)$$

Vidíme, že výpočet vyšších momentů se stává stále složitějším, takže je pak nutno k zjednodušení výpočtů užívati účelné symboliky, ale přes to i výsledky jsou pro praktické upotřebení velmi složité. Podstatné zjednodušení nastává pro případ, kdy základní soubor je nekonečného rozsahu.

(2,2) Momenty výběrových průměrů z nekonečného základního souboru. Necháme-li v rovnicích (3), (12), (13), (17), (20), (21), růsti $N \to \infty$ do nekonečna, a při tom zůstává r konečné, dostaneme pro výběrové momenty výrazy

$$\mu'_1 = \bar{x}, \qquad \mu_2 = \frac{1}{r}\,\mu(x, 2), \qquad (22)$$

$$\mu_3 = \frac{1}{r^2}\,\mu(x, 3), \quad \mu_4 = \frac{1}{r^3}\,[\mu(x, 4) + 3(r - 1)\,\mu^2(x, 2)], \qquad (23)$$

$$\mu_5 = \frac{1}{r^4}\,[\mu(x, 5) + 10(r - 1)\,\mu(x, 3)\,\mu(x, 2)], \ \ldots$$

odkud dostáváme pro směrodatnou odchylku výběrových průměrů

$$\sigma_P = \frac{\sigma(x)}{\sqrt{r}} \qquad (24)$$

a rovnice (22) a (23) můžeme snadno uvésti na tvar

$$\mu_2 = \frac{1}{r}\,\mu(x, 2),$$

$$\mu_3 = \frac{1}{r^2}\,\mu(x, 3), \tag{25}$$

$$\mu_4 - 3\mu_2^2 = \frac{1}{r^3}\,[\mu(x, 4) - 3\mu^2(x, 2)],$$

$$\mu_5 - 10\mu_3\mu_2 = \frac{1}{r^4}\,[\mu(x, 5) - 10\mu(x, 3)\,\mu(x, 2)],$$

odkud plynou pro charakteristiky směrodatné proměnné (I, str. 22 a 23) výrazy

$$\begin{aligned}
&\alpha_{P,2} = 1,\\
&\alpha_{P,3} = \frac{1}{r^{\frac{1}{2}}}\,\alpha(x, 3),\\
&\alpha_{P,4} - 3 = \frac{1}{r}\,[\alpha(x, 4) - 3],\\
&\alpha_{P,5} - 10\alpha_{P,3} = \frac{1}{r^{\frac{3}{2}}}\,[\alpha(x, 5) - 10\alpha(x, 3)],\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{26}$$

Budeme-li nyní psáti v (25)

$$\begin{aligned}
&\lambda_2 = \mu_2, && \lambda(x, 2) = \mu(x, 2),\\
&\lambda_3 = \mu_3, && \lambda(x, 3) = \mu(x, 3),\\
&\lambda_4 = \mu_4 - 3\mu_2^2, && \lambda(x, 4) = \mu(x, 4) - 3\mu^2(x, 2),\\
&\lambda_5 = \mu_5 - 10\mu_3\mu_2, && \lambda(x, 5) = \mu(x, 5) - 10\mu(x, 3)\,\mu(x, 2),
\end{aligned} \tag{27}$$

vidíme, že podle rovnic (25) je rozdělení výběrových průměrů z nekonečného základního souboru charakterisováno jednoduchým vztahem mezi λ-funkcemi

$$\lambda_n = \frac{1}{r^{n-1}}\,\lambda(x, n). \tag{28}$$

Jinou cestou objevil Thiele důležitost těchto λ-funkcí, které tolik přispěly k rozvoji teorie matematické statistiky a nazývají se Thieleho semiinvarianty.

Dále přicházíme k semiinvariantům směrodatné proměnné čili standardisovaným semiinvariantům Thieleho, píšeme-li v (26)

$$\begin{aligned} \gamma_3 &= \alpha_{P,3}, & \gamma(x, 3) &= \alpha(x, 3), \\ \gamma_4 &= \alpha_{P,4} - 3, & \gamma(x, 4) &= \alpha(x, 4) - 3, \\ \gamma_5 &= \alpha_{P,5} - 10_{P,3}, & \gamma(x, 5) &= \alpha(x, 5) - 10\alpha(x, 3), \end{aligned} \tag{29}$$

mezi nimiž platí vztah

$$\gamma_n = \frac{1}{r^{\frac{n}{2}-1}} \gamma(n, x). \tag{30}$$

Necháme-li zde růsti rozsah výběru $r \to \infty$ do nekonečna, bude $\lim_{r\to\infty} \gamma_n = 0$, z čehož podle (26) plyne

$$\alpha_{P,3} = 0, \quad \alpha_{P,4} = 3, \quad \alpha_{P,5} = 0, \ \ldots$$

a dále bychom dostali obecně

$$\alpha_{P,2n} = \frac{(2n)!}{2^n(n!)}, \quad \alpha_{P,2n+1} = 0,$$

což jsou momenty normálního rozdělení četností [I, (47), (48)].

Z toho vidíme, že pro velká r jsou momenty tohoto rozdělení četností shodné s momenty normálního rozdělení. To znamená, že bereme-li velké výběry z nekonečného základního souboru, můžeme očekávati, že rozdělení výběrových průměrů se těsně přiblíží normálnímu.

(2,2,l) Příklad. 1. Stanovme momenty výběrových průměrů ze základního souboru nekonečného, jedná-li se o alternativní znak s četností p. Označíme pozorovaný znak jedničkou a jeho nepřítomnost nulou, takže hodnota znaku $x_1 = 1$ má relativní četnost p, $x_2 = 0$ má relativní četnost $q = 1 - p$. Průměr $\bar{x} = p$. Bude tudíž n-tý moment vzhledem k průměru

$$\mu(x, n) = \sum (x_i - \bar{x})^n f_i$$

vyjádřen

$$\begin{aligned} \mu(x, n) &= (1 - p)^n p + (-p)^n (1 - p) = \\ &= pq [q^{n-1} + (-1)^n p^{n-1}]. \end{aligned}$$

Výběry rozsahu r z tohoto základního souboru mají vzhledem k počtu prvků s pozorovaným znakem rozdělení binomické $\binom{r}{x} p^x q^{r-x}$.

Momenty výběrových průměrů pro proměnnou x pak stanovíme podle rovnic (22) a (23); dostáváme

$$\mu'_1 = p, \qquad \mu_2 = \frac{1}{r} pq,$$

$$\mu_3 = \frac{1}{r^2} pq\,(q^2 - p^2),$$

$$\mu_4 = \frac{1}{r^3} pq\,(q^3 + p^3) + \frac{3}{r^3} p^2q^2\,(r - 1).$$

Příklad 2. Vypočítejme momenty výběrových průměrů ze základního souboru:

a) S rozdělením normálním.

Vzhledem k rovnicím [I, (48), str. 80] můžeme podle rovnic (22) a (23) psáti

$$\mu'_1 = \bar{x}, \qquad \mu_2 = \frac{1}{r} \sigma^2(x), \qquad \mu_3 = 0,$$

$$\mu_4 = \frac{1}{r^3} [3\sigma^4(x) + 3(r-1)\,\sigma^4(x)] = \frac{3}{r^2} \sigma^4(x).$$

Z rovnic (27) bychom se přesvědčili, že všechny semiinvarianty vyššího stupně než druhého jsou identicky rovny nule.

b) S rozdělením podle Pearsonovy křivky typu III.

Typ III ze systému křivek Pearsonových (Janko [1], str. 42), lze pomocí momentů směrodatné proměnné uvésti na tvar

$$y = y_0 \left(1 + \frac{\alpha(x,3)}{2}\, t\right)^{\frac{4}{\alpha^2(x,3)} - 1} e^{-\frac{2}{\alpha(x,3)} t}.$$

Počátek souřadnic je v průměru a první momenty směrodatné proměnné vzhledem k němu jsou $\alpha(x, 0) = 1, \alpha(x, 1) = 0$, $\alpha(x, 2) = 1$.

Pro další momenty pak je odvozena rekurentní formule

$$\alpha(x, n+1) = n\left[\alpha(x, n-1) + \frac{\alpha(x, 3)\,\alpha(x, n)}{2}\right]. \quad (31)$$

Z toho tedy plyne vyjádření dalších momentů pomocí $\alpha(x, 3)$, takže

$$\alpha(x, 4) = 3\left(1 + \frac{\alpha^2(x, 3)}{2}\right),$$

$$\alpha(x, 5) = 2\alpha(x, 3)\left[5 + \frac{3\alpha^2(x, 3)}{2}\right], \ldots$$

a semiinvarianty směrodatné proměnné jsou podle (29)

$$\gamma(x, 4) = \frac{3\alpha^2(x, 3)}{2} = \left(\frac{\alpha(x, 3)}{2}\right)^2 3!,$$

$$\gamma(x, 5) = \frac{2.3\alpha^3(x, 3)}{2} = \left(\frac{\alpha(x, 3)}{2}\right)^3 4!. \quad (32)$$

Semiinvarianty γ_4 a γ_5 rozdělení průměrů jsou pak dány podle rovnice (30).

Vypočítáme však momenty směrodatné proměnné podle rovnic (26) za použití rekurentního vztahu (31). Tak dostaneme

$$\alpha_{P,2} = 1, \qquad \alpha_{P,3} = \frac{1}{r^{\frac{1}{2}}}\,\alpha(x, 3), \qquad \alpha_{P,4} = \frac{3}{r}\left[r + \frac{\alpha^2(x, 3)}{2}\right],$$

$$\alpha_{P,5} = \frac{2}{r^{\frac{3}{2}}}\,\alpha(x, 3)\,[5r + \tfrac{3}{2}\alpha^2(x, 3)],$$

jako momenty rozdělení průměrů, jestliže se berou výběry rozsahu r z nekonečného základního souboru Pearsonova typu III. Pro čtvrtý a pátý moment se můžeme přesvědčiti, že vyhovují podmínce

$$\alpha_{P,n+1} = n\,(\alpha_{P,n-1} + \frac{\alpha_{P,3}}{2}\,\alpha_{P,n});$$

poněvadž i další momenty této podmínce vyhovují, vidíme, že rozdělení průměrů je tu dáno také Pearsonovým typem III.

(2,3) Momenty rozdělení četností výběrových rozptylů. Rozdělení četností průměrů bylo známo již v minulém století. Je však jednou z nejvýznamnějších složek pokroku tohoto století ve statistice, že se začalo studovati rozdělení četností rozptylů.

Průměr. Než přistoupíme k výpočtu těchto charakteristik, je dobře zvláště si uvědomiti, že se jedná o druhé momenty kolem průměru každého jednotlivého výběru. Tak rozptyl výběru čili druhý moment kolem průměru tohoto výběru je podle definice

$$\sigma_{x,i}^2 = \frac{1}{r} \sum^{r,i} (x_j - \bar{x}_i)^2, \tag{33}$$

kde se součet vztahuje na všech r prvků i-tého výběru.

Tento výraz upravíme, abychom do něho zavedli jen hodhoty znaku

$$\sum^{r,i} (x_j - \bar{x}_i)^2 = \sum^{r,i} x_j^2 - 2\bar{x}_i \sum^{r,i} x_j + r\bar{x}_i^2 = \sum^{r,i} x_j^2 - r\bar{x}_i^2,$$

kde jsme psali $r\bar{x}_i = \sum^{r,i} x_j$, takže bude dále

$$\sum^{r,i} x_j^2 - \frac{1}{r}\left(\sum^{r,i} x_j\right)^2 = \frac{1}{r}\left[r \sum^{r,i} x_j^2 - \sum^{r,i} x_j^2 - 2 \sum_{j,k}^{r,i} x_j x_k\right]$$

a tedy

$$\sigma_{x,i}^2 = \frac{1}{r^2}\left[(r-1) \sum^{r,i} x_j^2 - 2 \sum_{j,k}^{r,i} x_j x_k\right].$$

Sečteme-li všech ν výběrových rozptylů a dělíme jejich počtem, dostaneme průměr $\overline{\sigma^2}$ podobnými úvahami jako při

odvození rovnice (7)

$$\bar{\sigma}^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}\sigma_{x,i}{}^2 = \frac{1}{\nu}\cdot\frac{1}{r^2}\left[\frac{r\nu}{N}(r-1)\sum_{j=1}^{N}x_j{}^2 - 2\frac{\binom{r}{2}\nu}{\binom{N}{2}}\sum_{j,k}^{N}x_j x_k\right],$$

takže vzhledem k rovnicím (7) a (9) bude

$$\bar{\sigma}^2 = \frac{1}{r^2}N\mu(x,2)\left[(r-1)\vartheta_1 + \vartheta_2\right] = \frac{1}{r^2}N^2\vartheta_2\mu(x,2), \qquad (34)$$

což lze psáti

$$\bar{\sigma}^2 = \frac{r-1}{r}\frac{N}{N-1}\mu(x,2). \qquad (35)$$

Průměrný rozptyl výběrový je menší než rozptyl základního souboru.

Z rovnic (12) a (35) je zřejmo, že platí vztah

$$\mu_2 + \bar{\sigma}^2 = \mu(x,2). \qquad (36)$$

Rozptyl rozdělení četností výběrových rozptylů. Pro zjednodušení výpočtu druhého momentu $\mu(\sigma_{x,i}{}^2, 2)$ výběrových rozptylů kolem jejich průměru (35) by bylo třeba užíti vhodného symbolického počtu, jak jsme se již dříve zmínili. Pro náš účel se však zde spokojíme sdělením výsledku

$$\mu(\sigma_{x,i}{}^2, 2) = \frac{1}{r^4}N\mu(x,4)\left[(r-1)^2\vartheta_1 - (r^2-6r+7)\vartheta_2 - \right.$$

$$\left. -4(r-3)\vartheta_3 - 6\vartheta_4\right] + \frac{1}{r^4}N^2\mu^2(x,2)\left[(r^2-2r+3)\vartheta_2 + \right.$$

$$\left. + 2(r-3)\vartheta_3 + 3\vartheta_4 - N^2\vartheta_2{}^2\right]$$

čili

$$\mu(\sigma_{x,i}{}^2, 2) = \frac{N(r-1)(N-r)}{r^3(N-1)^2(N-2)(N-3)}$$

$$\{(N-1)(rN-N-r-1)\mu(x,4) +$$

$$+ [(3-r)N^2 - 6N + 3r + 3]\mu^2(x,2)\}. \qquad (37)$$

Druhou odmocninou tohoto výrazu je pak dána směrodatná odchylka výběrových rozptylů $\sigma(\sigma_{x,i}^2)$.

Výrazy pro další momenty jsou ještě mnohem rozsáhlejší a nebudeme je uváděti, ježto také praktická cena jejich je tím omezena.

(2,4) Momenty výběrových rozptylů z nekonečného základního souboru. Průměr výběrových rozptylů je dán rovnicí (35), v níž necháme nyní růsti rozsah $N \to \infty$ do nekonečna, takže potom

$$\overline{\sigma}^2 = \frac{r-1}{r}\mu(x, 2). \tag{38}$$

Rozptyl jejich dostaneme obdobně z rovnice (37), z níž pro $N \to \infty$ vyplývá

$$\mu(\sigma_{x,i}^2, 2) = \frac{r-1}{r^3}[(r-1)\mu(x, 4) - (r-3)\mu^2(x, 2)] =$$

$$= \frac{r-1}{r^3}\sigma^4(x)[(r-1)\alpha(x, 4) - (r-3)] \tag{39}$$

a směrodatnou odchylku můžeme tedy psáti ve tvaru

$$\sigma(\sigma_{x,i}^2) = \frac{\sigma^2(x)}{r}\sqrt{\frac{r-1}{r}[(r-1)\alpha(x, 4) - r + 3]}. \tag{40}$$

Uvedeme ještě třetí moment výběrových rozptylů $\mu(\sigma_{x,i}^2, 3)$ z nekonečného základního souboru.

$$\mu(\sigma_{x,i}^2, 3) = \frac{r-1}{r^5}\sigma^6(x)[(r-1)^2\alpha(x, 6) -$$

$$- 3(r-1)(r-5)\alpha(x, 4) - 2(3r^2 - 6r + 5)\alpha^2(x, 3) +$$
$$+ 2(r^2 - 12r + 15)].$$

Velmi značné zjednodušení nastává pro případ základního souboru s normálním rozdělením četností. Výpočtem semiinvariantů lze prokázat, že rozdělení výběrových rozptylů kolem průměru je Pearsonova křivka typu III.

Charakteristiky rozdělení četností třetích a čtvrtých výběrových momentů.

Uvedeme jen výsledky pro průměr a rozptyl třetích výběrových momentů kolem průměru z nekonečného základního souboru $N \to \infty$.

Průměr třetích výběrových momentů

$$\mu'(\mu_{3,i},\, 1) = \frac{(r-1)\,(r-2)}{r^2}\,\mu(x,\, 3) =$$

$$= \frac{r-1)\,(r-2)}{r^2}\,\sigma^3(x)\,\alpha(x,\, 3)$$

rozptyl

$$\sigma^2(\mu_{3,i}) = \frac{(r-1)\,(r-2)}{r^5}\,\sigma^6(x)\,[(r-1)\,(r-2)\,\alpha(x,\, 6) -$$

$$- 3(r-2)\,(2r-5)\,\alpha(x,\, 4) - (r-2)\,(r-10)\,\alpha^2(x,\, 3) +$$

$$+ 3\,(3r^2 - 12r + 20)].$$

Průměr rozdělení čtvrtých momentů

$$\mu'(\mu_{4,i},\, 1) = \frac{r-1}{r^3}\,[(r^2 - 3r + 3)\,\mu(x,\, 4) +$$

$$+ 3\,(2r-3)\,\mu^2(x,\, 2)].$$

Odvozením charakteristik výběrových pomocí parametrů základního souboru jsme získali možnost řešiti v případech, kde známe rozdělení četností dotyčné charakteristiky otázku, jaká je pravděpodobnost, že na př. průměr náhodného výběru bude v daných mezích. Tak víme, že bude v intervalu $\pm 2\sigma_P$ s pravděpodobností 0,955, nebo v intervalu $\pm 3\sigma_P$ s pravděpodobností 0,997, když se jedná o velký výběr z nekonečného základního souboru, t. j. výběr, který byl vzat tak, že se každý prvek po zjištění příslušné hodnoty znaku vrátil zpět do základního souboru čili jeho složení zůstávalo nezměněno. V tom případě totiž můžeme považovat rozdělení četností průměrů za normální, jak jsme si odvodili. Obdobně můžeme postupovati, známe-li rozdělení

četností výběrových rozptylů, nebo můžeme-li o něm předpokládati, že je vyjádřeno na př. Pearsonovou křivkou typu III. Integrály této křivky jsou dány buď Pearsonovými tabulkami neúplné Γ-funkce nebo výhodněji pro směrodatnou proměnnou tabulkami Salvosovými.

Budeme potřebovati hlavně k praktickému použití směrodatné odchylky některých charakteristik, které zde ještě uvedeme, ale již bez odvození.

Směrodatná odchylka rozdělení směrodatných odchylek výběrových z normálního základního souboru je

$$\sigma_\sigma = \frac{\sigma(x)}{\sqrt{2r}} \tag{41}$$

a nesmí se jí ovšem užívati bez tohoto zřetele k formě rozdělení četností základního souboru. V obecném případě by totiž bylo nutno pracovati s výrazem

$$\sigma_\sigma = \frac{\sigma(x)}{\sqrt{2r}} \left(1 + \frac{\frac{\mu(x,4)}{\mu^2(x,2)} - 3}{2}\right)^{\frac{1}{2}}. \tag{42}$$

Je-li exces $\left[\frac{\mu(x,4)}{\mu^2(x,2)} - 3\right]$ malý, pak odmocnina výrazu v kulaté závorce je přibližně rovna $1 + \frac{1}{4}\left[\frac{\mu(x,4)}{\mu^2(x,2)} - 3\right]$ a toto číslo se liší od jednotky o více než 5 procent, je-li $\frac{\mu(x,4)}{\mu^2(x,2)}$ menší než 2,8 nebo větší než 3,2.

Dále uvedeme ještě směrodatnou odchylku rozdělení četností šikmosti ve výběrech z normálního základního souboru, která je $\sqrt{\frac{6}{r}}$ a excesu, která je $2\sqrt{\frac{24}{r}}$.

(2,4,1) Příklad 1. Základní soubor má rozsah 1000 prvků; průměr měřeného znaku je 140 jednotek. Jaký průměr mů-

žeme očekávat, vezmeme-li z něho náhodný výběr rozsahu $r = 500$, nebo $r = 100$. Vyložte, zda by k odpovědi pomohla znalost směrodatné odchylky rozdělení četnosti výběrových průměrů. Odpověď: můžeme očekávat průměr blízký hodnotě $\bar{x} = 140$ jednotek a to v prvním případě bližší než v druhém. Kdybychom znali směrodatnou odchylku σ_P, která by byla v každém z obou případů jiná, mohli bychom říci, že pravděpodobná odchylka od hodnoty $\bar{x} = 140$ bude $\pm 0{,}6745\,\sigma_P$.

Příklad 2. Základní soubor je rozsahu $N = 1000$ prvků, průměr hodnot pozorovaného znaku je 67,6 cm a směrodatná odchylka 2,5 cm. Jaká je směrodatná odchylka všech možných výběrových průměrů, je-li rozsah každého výběru 200, resp. 500 nebo 800 prvků? Použijeme-li výrazu (13) dostaneme $\sigma_P = 0{,}158;\ 0{,}08;\ 0{,}04$. Znázorněte graficky výraz (13) pro svrchu uvedené N a $\sigma(x)$ a proveďte diskusi. Jak velký musí býti rozsah výběrů r, aby σ_P bylo menší než $\frac{1}{2}\sigma(x)$ resp. $\frac{1}{10}\sigma(x)$?

$$\sigma_P = \frac{\sigma(x)}{\sqrt{r\,\dfrac{N-1}{N-r}}} < \frac{\sigma(x)}{2}$$ znamená, že $r\,\dfrac{N-1}{N-r} > 4$ čili $r > \dfrac{4N}{N+3} \doteq 3{,}99$. V druhém případě $r\,\dfrac{N-1}{N-3} > 100$ čili $r > \dfrac{100N}{N+99} \doteq 90{,}99$.

Výraz pro σ_P dává pro některá r tyto hodnoty σ_P:

r	1	10	100	200	500	800	N
σ_P	$\sigma(x) = 2{,}5$	0,79	0,24	0,14	0,08	0,04	0

Z tohoto přehledu a jeho grafického znázornění je patrno, že σ_P je klesající funkcí r a v našem případě klesá od hodnoty směrodatné odchylky základního souboru $\sigma(x) = 2{,}5$ tak, že pro $r = 4$ má hodnotu menší než její polovina, pro $r = 91$

menší než desítina a dospěje k nule, je-li rozsah výběru totožný s rozsahem základního souboru.

Příklad 3. Známe z tab. 1 (str. 10) rozdělení četností základního souboru rozsahu $N = 1001$ podle pozorovaného znaku. Průměr je $\bar{x} = 7{,}000$, směrodatná odchylka $\sigma(x) = 2{,}002$. Šikmost nemusíme bráti v úvahu, neboť je rovna nule. Jaká je pravděpodobnost, že ve výběru rozsahu $r = 200$ bude průměr větší než 7,124, nebo větší než 7,174, nebo 7,500?

$$\sigma_P = 2{,}002 \sqrt{\frac{801}{200 \cdot 1000}} = \frac{2{,}002}{15{,}8015} = 0{,}127,$$

$$t_1 = \frac{7{,}124 - 7{,}000}{0{,}127} = 0{,}976, \qquad t_2 = \frac{0{,}174}{0{,}127} = 1{,}370,$$

$$t_3 = \frac{0{,}500}{0{,}127} = 3{,}937.$$

Vzhledem k povaze základního souboru můžeme použíti tabulky integrálu Laplace-Gaussova ([1], str. 38), takže

$$p = 0{,}5 - \frac{\alpha(t)}{2}$$

a dostáváme

$$p_1 = 0{,}165$$
$$p_2 = 0{,}085$$
$$p_3 = 0{,}00004.$$

Ve skutečnosti bylo podle sloupce (7) tabulky 1. větších průměrů výběrových než 7,124 celkem 70 ze 400, tedy $\sum f = 0{,}175$; větších než 7,174 bylo 30, takže $\sum f = 0{,}075$ a větší než 7,500 se nevyskytl.

Pozorovaná směrodatná odchylka byla $\sigma'_P = 0{,}121$.

Příklad 4. Základní soubor rozsahu $N = 10\,000$ prvků má průměr $\bar{x} = 69{,}7$ jednotek a $\sigma(x) = 7{,}4$. Jaký rozsah musí míti výběr, aby směrodatná odchylka průměrů byla menší než 2, 1, 0,5 jednotky?

Obecně má býti splněna nerovnost

$$\sigma_P = \frac{\sigma(x)}{\sqrt{r\,\frac{N-1}{N-r}}} < k,$$

odkud $$\sigma^2(x)\,(N-r) < k^2 r(N-1),$$

$$r > \frac{\sigma^2(x)\,N}{k^2(N-1)+\sigma^2(x)}.$$

Pro $k = 2$ dostáváme $r > \frac{7{,}4^2 \cdot 10\,000}{4{,}9999 + 7{,}4^2} = 13{,}67$ a podobně pro $k = 1$ je $r > 54{,}47$ a pro $k = 0{,}5$ je $r > 214{,}37$. Poněvadž rozsah výběru je číslo celé, vidíme, že musí býti v prvém případě $r \geqq 14$, v druhém případě $r > 54$ a ve třetím $r > 214$.

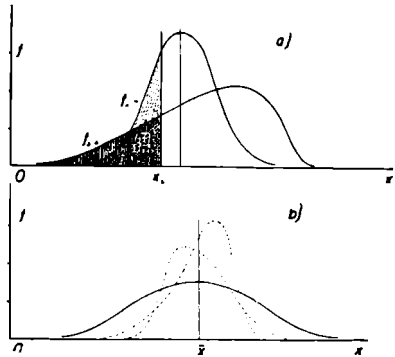
Příklad 5. Je-li v základním souboru nekonečného rozsahu šikmost 1,1, jak musí býti velký náhodný výběr, aby šikmost rozdělení všech možných výběrových průměrů byla menší než 0,1? Nakreslete křivku $\frac{1{,}1}{\sqrt{r}}$ a proveďte rozbor.

Podle rovnic (26) $\alpha_{P,3} = \frac{1}{r^{\frac{1}{2}}}\,\alpha(x, 3)$ a podle úlohy má býti splněna podmínka $0{,}1 < \frac{\alpha(x,3)}{\sqrt{r}}$, tedy $0{,}1 < \frac{1{,}1}{\sqrt{r}}$, takže $r < \frac{1{,}21}{0{,}01} = 121$. Křivka s počátku rychle klesá od hodnoty 1,1 pro $r = 1$ a blíží se asymptoticky nule.

(2,5) Podstatná informace v parametrech a charakteristikách. Než přikročíme k výkladu inversního úkolu odhadování parametrů základního souboru podle zjištěných charakteristik výběrových, osvětlíme si otázku, jak mnoho informace podávají jednotlivé charakteristiky o souboru, z něhož byly stanoveny. Budeme předpokládati k tomuto účelu, že

celá informace o určitém souboru vzhledem k pozorovanému znaku je obsažena v původní množině čísel $x_1, x_2, \ldots, x_N$ seřazených vzestupně podle velikosti (uspořádaně), tedy neseskupených do tříd (viz na př. I. díl, str. 19). Mnoho z této celé informace je obsaženo v několika málo charakteristikách. Máme tedy za úkol udati, v jakých charakteristikách musíme podati zhuštěně informaci, aby z ní vyplývalo těsné přiblížení k uspořádané posloupnosti čísel x_i, což znamená jinými slovy, aby bylo co možná s nejlepším přiblížením vyjádřeno procento prvků z celého počtu N, které spadají do určitého intervalu (x_i, x_{i+k}) hodnot znaku.

Zhuštěním informace o pozorovaném materiálu pomocí skupinového rozdělení četností jsme se zabývali v I. dílu (str. 29). Představme si, že uvedeme jen jednu hodnotu relativní četnosti p_k, která udává zlomek z celého počtu N pozorovaných hodnot x_i, které jsou menší než x_k.



Obr. 2. Množství informace obsažené jen a) v relativní četnosti, b) v průměru.

Je zřejmo, že tím dáváme jen velmi malou část celé informace, neboť tuto hodnotu p_k mohou vykazovati rozdělení četností, která se od sebe zcela liší (viz v obr. 2 plochy f_1 a f_2).

Také velmi malou část celé informace podáváme, uvedeme-li jen průměr $\bar{x}$ a rozsah pozorovaného souboru N. Touž hodnotu $\bar{x}$ mohou totiž míti úplně odlišná rozdělení četností (obr. 2b).

Vidíme tudíž, že jedna charakteristika nemůže sama podati mnoho z celé informace obsažené v původní posloup-

nosti. Teprve uvedením dvou nebo tří charakteristik můžeme dosáhnouti značně úplného vystižení rozdělení četností.

Dva parametry $\bar{x}, \sigma(x)$ nám udávají spolu s rozsahem N, že víc než $1 - \frac{1}{\tau^2}$ z celkového počtu r prvků souboru je v mezích $\bar{x} \pm \tau\, \sigma(x)$ (kde $\tau \geqq 1$) (viz I. díl, str. 72) a to bez jakékoliv výhrady a bez ohledů na tvar rozdělení četností. Můžeme-li podle některých okolností, na př. podle původu materiálu, souditi na tvar rozdělení četností, dosáhneme výstižnějšího odhadu počtu prvků v uvedeném intervalu. Na př. pro normální rozdělení četností jsou dotyčné zlomky uvedeny v procentech v I. díle, str. 83.

Tak se vyskytuje normální rozdělení často při kontrole hromadné výroby nějakého předmětu, kde prvky pozorovaného souboru byly vyrobeny za týchž podstatných podmínek a rovněž pozorování znaku vyplynulo z měření za týchž podstatných podmínek. Pak se obyčejně říká, že data byla získána za kontrolovaných podmínek. V takových případech bývá možno pro většinu praktických účelů předpokládati, že rozdělení četností je normální nebo mírně nesymetrické.

Uvádíme-li kromě průměru a směrodatné odchylky ještě šikmost $\alpha(x, 3)$, přispíváme málo k vystižení rozdělení četností v symetrických intervalech kolem průměru a tedy k podání celé informace obsažené v pozorovaných datech. V takovém případě je třeba pomáhat si plochami křivky nesymetrické a v intervalech, jejichž hranice nejsou souměrně rozloženy nad a pod průměrem. Praktikové přikládají tomuto postupu význam tehdy, je-li rozsah pozorovaného souboru větší než $N = 250$.

Zkušenosti s rozděleními četnosti pro fysikální vlastnosti materiálu a výrobků v továrnách vyráběných hromadně potvrzují, že se vystačí se způsoby právě uvedenými pro hrubé odhady pozorovaných procent nějakého rozdělení četností aspoň pro symetrické intervaly kolem průměru, ježto jsou to zpravidla rozdělení s jedním maximem. Není ovšem možno očekávati, že bychom s nimi vystačili v případě bi-

modálního rozdělení, které je výsledkem dvou různých množství podmínek, leč že bychom mohli rozštěpiti soubor ve dvě skupiny dat, z nichž každá by měla rozdělení jednovrcholové.

Záleží na účelu, k němuž hodláme použíti pozorovaných dat, abychom mohli rozhodnouti, které charakteristiky mají býti uvedeny v konkretním případě. Uvádíme je proto, abychom z nich mohli odvoditi hledané závěry nebo, aby jich mohl užíti statistický spotřebitel. Chceme tedy, aby obsahovaly podstatnou informaci. Co tvoří podstatnou informaci v určitém případě závisí na povaze otázek, které máme zodpověděti a na povaze hypotés, které chceme učiniti na základě informace z dat k tomu cíli dosažitelných. Říkáme, že nějaká skupina charakteristik obsahuje podstatnou informaci danou pozorovanými daty, když můžeme pomocí těchto charakteristik odpověděti na položené otázky tak, že by další rozbor dat naše odpovědi prakticky nepozměnil.

Praxe pak dospívá k názoru, že při studiu jednoho znaku, vzniklého za týchž podstatných podmínek na prvcích souboru, obsahuje průměr, směrodatná odchylka a rozsah souboru podstatnou informaci ve většině případů. Bývá tomu tak, když se zajímáme o průměrnou jakost materiálu a variabilitu průměrů postupných výběrů, nebo když srovnáváme tyto vlastnosti určitého materiálu s jiným. Kdybychom potřebovali k zodpovědění položených otázek znáti procenta těch prvků z celkového počtu pozorovaných prvků, kde hodnoty studovaného znaku jsou větší nebo menší než určitá daná hodnota, pak může býti podstatná informace obsažena v tabulce třídního rozdělení četností.

Totéž, co zde bylo řečeno o parametrech, tedy pro základní soubor, platí o charakteristikách pro náhodný výběr.

(3) Náhodné výběry z neznámého základního souboru.

Zabývali jsme se úkolem: najíti pravděpodobnost, že charakteristika náhodného výběru bude v daných mezích, když známe parametry základního souboru.

Prakticky důležitější je ve statistice úkol obrácený: najíti pravděpodobnost, že parametry základního souboru se neliší od výběrových charakteristik určitého pozorovaného výběru o více než o daný zlomek jejich. Otázka může býti také jinak položena. Je známa z pozorování charakteristika výběrová; máme udati meze, v nichž musí býti hledaný parametr základního souboru s určitou napřed danou pravděpodobností.

Nebo konkretně, ale méně přesně můžeme říci, že tato otázka znamená

a) do jaké míry se shoduje průměr výběru s průměrem základního souboru,

b) jak dobře se shoduje rozptyl a vyšší momenty výběru s příslušnými v základním souboru,

c) jak těsně blízko je křivka rozdělení četností výběru u křivky rozdělení četností základního souboru.

Vypočítáme-li z velkého výběru nějakou charakteristiku, třeba průměr $\bar{x}_i$, přisuzujeme jí obyčejně velkou přesnost. Můžeme-li vzíti ze základního souboru nekonečného rozsahu více výběrů a pro každý vypočítati průměr, budou se zpravidla rozdíly mezi nimi zmenšovat, když rozsah výběrů poroste. Podle věty Bienaymé-Čebyševovy ([1], str. 70) usuzujeme, že rozdíl $|\bar{x}_i - \bar{x}|$ průměru pozorovaných výběrů a očekávaného průměru v základním souboru bude menší než libovolné kladné číslo $\eta = \tau\sigma_P$ s pravděpodobností

$$P = 1 - P_\tau > 1 - \frac{\sigma_P{}^2}{\eta^2}.$$

Poněvadž $\sigma_P{}^2 = \frac{1}{r}\sigma^2(x)$ — podle (22) — a rozptyl v základním souboru předpokládáme jako veličinu kladnou, blíží se $P > 1 - \frac{\sigma^2(x)}{r\eta^2}$ s rostoucím rozsahem výběru r k jednotce čili pravděpodobnost, že se průměr výběrový $\bar{x}_i$ liší libovolně málo od průměru v základním souboru $\bar{x}$ může se libovolně přiblížiti jednotce. Za těchto okolností říkáme, že

průměr pozorovaných dat (platí to také pro jiné charakteristiky výběrové) konverguje stochasticky k své očekávané hodnotě, t. j. k hodnotě v základním souboru (k příslušnému parametru).

(3, I) Charakteristiky konsistentní, efficientní, sufficientní. Můžeme to vyjádřiti také tak, že nejvhodnějším odhadem parametru je charakteristika vypočítaná z náhodného výběru, jejíž očekávaná (průměrná) hodnota dává hledaný výsledek, t. j. hodnotu parametru. Takové charakteristiky se nazývají konsistentními čili souhlasnými odhady parametru.

Obecně existují různé charakteristiky, které mohou míti touž očekávanou hodnotu: na př. $\mathfrak{E}(x_i) = \bar{x}$ a také

$$\mathfrak{E}\left(\frac{1}{r}\sum_{i=1}^{r} x_i\right) = \bar{x}.$$

Zavádí se proto další kriterium, které dovoluje mezi několika konsistentními charakteristikami téhož parametru vybrati poměrně nejvhodnější. Za takovou se považuje ta charakteristika z konsistentních, která má normální rozdělení četností pro výběry velkého rozsahu r a poměrně nejmenší rozptyl: nazývá se efficientní čili vydatná nebo výstižná. Pro objasnění uvažujeme prvky $x_1, x_2, \ldots, x_j, \ldots, x_r$ výběru rozsahu r. Touž očekávanou hodnotu $\bar{x}$ budou míti výrazy

$$x_1, \quad \frac{x_1 + x_2}{2}, \quad \ldots, \quad \frac{1}{r-1}\sum_{j=1}^{r-1} x_j, \quad \frac{1}{r}\sum_{r=1}^{r} x_j$$

a jsou tedy konsistentní. Jejich rozptyly vzhledem k očekávané hodnotě však jsou

$$\mathfrak{E}(x_1 - \bar{x})^2 = \mu(x, 2),$$

$$\mathfrak{E}\left(\frac{x_1 + x_2}{2} - \bar{x}\right)^2 = \left(1 - \frac{1}{N-1}\right)\frac{\mu(x, 2)}{2},$$

.....

$$\mathfrak{E}\left(\frac{1}{r-1}\sum_{j=1}^{r-1} x_j - \bar{x}\right)^2 = \left(1 - \frac{r-2}{N-1}\right)\frac{\mu(x,2)}{r-1},$$

$$\mathfrak{E}\left(\frac{1}{r}\sum_{j=1}^{r} x_j - \bar{x}\right)^2 = \left(1 - \frac{r-1}{N-1}\right)\frac{\mu(x,2)}{r},$$

takže výběrový průměr $\frac{1}{r}\sum_{j=1}^{r} x_j$ považovaný za odhad parametru $\bar{x}$ má poměrně nejmenší rozptyl. Vzhledem k tomu pak, že rozdělení četností spěje při rostoucím r k normálnímu, je efficientní čili vydatný. Vydatnost ostatních srovnávaných výrazů se měří obráceným poměrem jejich rozptylů k rozptylu nejvydatnějšího. Tak bude na př. vydatnost x_1 u srovnání s průměrem $\frac{1}{r}\sum_{j=1}^{r} x_j$ vyjádřena zlomkem

$$\left(1 - \frac{r-1}{N-1}\right)\frac{\mu(x,2)}{r} : \mu(x,2) = \frac{N-r}{(N-1)\,r}.$$

Lze také říci, že v tomto smyslu obsahuje výběrový průměr myslitelně nejúplnější informaci o parametru základního souboru u srovnání s ostatními výrazy, které mohou býti z dat výběru rozsahu r počítány.

Nemůže tudíž výpočet těchto výrazů jako $\frac{1}{r-1}\sum_{j=1}^{r-1} x_j, \ldots$ přispěti ničím novým k informaci podávané průměrem. Charakteristiky tohoto druhu mají také své pojmenování jako sufficientní čili vyčerpávající.

Je zřejmo, že můžeme vydatnost charakteristik vyjadřovat také v procentech té nejvydatnější. Tak můžeme porovnáním rozptylů průměru a mediánu zjistiti ([1], str. 184), že vydatnost mediánu klesá při rostoucím rozsahu výběru z normálního základního souboru na 80% při $r = 4$ a pak dále asymptoticky na 63%. Z toho na př. vyplývá, že medián obsahuje jen asi 63% té informace, kterou podává průměr a to z výběru rozsahu již asi $r = 25$. Objeví nám tedy průměr

z výběru o rozsahu $r=63$ změnu v poloze základního souboru stejně dobře jako medián z výběru rozsahu teprve $r = 100$. Podobně můžeme srovnati dvě charakteristiky rozptylu a to se směrodatnou odchylkou σ průměrnou odchylku ϑ. Vydatnost směrodatné odchylky je v případě normálního rozdělení četností o 12% vyšší než průměrné odchylky. Je tudíž výhodnější vynaložiti trochu více práce výpočtu výrazu $\frac{N-1}{N} \cdot \frac{r}{r-1} \overline{\sigma}^2$ — vzhledem k rovnici (35) — než zvětšiti počet pozorování asi o 14%, což by bylo nutné, kdybychom chtěli dosáhnouti pomocí průměrné odchylky té přesnosti jako při směrodatné odchylce, které se proto užívá téměř výhradně pro přesnější výpočty statistické.

(3,2) Odhad průměru. V praxi statistické musíme zpravidla udati určitou hodnotu charakteristiky vypočítanou z náhodného výběru, kterou lze považovati za nejvhodnější odhad velikosti neznámého parametru. Jedná-li se konkretně o průměr, uvědomíme si, že průměr všech možných výběrových průměrů byl týž jako průměr základního souboru; to jsme viděli v případě, kdy základní soubor byl znám (str. 14). Také jsme si vyložili (str. 11), že průměr výběru velkého rozsahu se mnoho neliší od průměru základního souboru. Ježto v praxi nebereme nikdy všechny možné výběry, nýbrž jeden nebo dva výběry dostačujícího rozsahu, užijeme jednoho zjištěného průměru nebo průměru dvou výběrů jako odhadu průměru základního souboru. Tento výběrový průměr bude representovati tedy průměr neznámého základního souboru a bude považován za jemu blízký, ježto variační obor rozdělení četnosti všech možných výběrových průměrů je velmi malý [viz formuli (13) nebo příklad (2,4,1,3), str. 31]. Proto se přijímá výběrový průměr za dobrý odhad průměru základního souboru.

(3,3) Odhad směrodatné odchylky. Potřebujeme nyní směrodatnou odchylku všech možných výběrových průměrů čili směrodatnou odchylku jejich rozdělení četností, když zá-

kladní soubor neznáme. Tato směrodatná odchylka nám bude naznačovat jak je přesný průměr, který jsme dostali z jednoho nebo několika málo výběrů. V případě, kdy základní soubor byl znám, jsme nalezli směrodatnou odchylku výběrových průměrů (13) a vyjádřili ji pomocí známé směrodatné odchylky základního souboru.

Tohoto výrazu však nemůžeme užíti, když základní soubor neznáme a tedy směrodatná odchylka jeho také není známa. Stojíme tudíž před problémem odhadu tohoto parametru. Vezmeme tedy ku pomoci průměr všech možných rozptylů výběrových (35). Musíme si připomenout, že každý rozptyl výběrový je počítán vzhledem k svému výběrovému průměru; z těchto rozptylů se pak vzal průměr. V tom je podstatný rozdíl proti výpočtu rozptylu výběrových průměrů vzhledem k průměru základního souboru (12) resp. (13).

Průměr všech výběrových rozptylů tedy je

$$\overline{\sigma}^2 = \frac{r-1}{r} \frac{N}{N-1} \sigma^2(x). \tag{43}$$

Kdybychom mohli vzíti aspoň několik výběrů a našli průměr jejich rozptylů, byl by lepším odhadem veličiny $\overline{\sigma}^2$, než vezmeme-li jeden náhodný výběr o r prvcích, který může dáti hodnotu rozptylu, která není daleko od $\overline{\sigma}^2$, ale také nemusí. Rozhodně však budeme považovati veličinu $\overline{\sigma}^2$ za nejlepší odhad rozptylu kolem průměru kteréhokoliv náhodného výběru nebo průměru rozptylů několika výběrů. Poněvadž $\overline{\sigma}^2$ neznáme, budeme považovati obráceně za jeho odhad rozptyl výběrový $\sigma_{x,v}{}^2$ nějakého výběru dosti velkého rozsahu a položíme tedy

$$\sigma_{x,v}{}^2 = \frac{(r-1)N}{r(N-1)} \sigma^2(x). \tag{44}$$

Jestliže veškerá informace, kterou máme, je z výběru, musíme podle ní odhadnouti s dobrým přiblížením parametry základního souboru, zde tedy $\sigma(x)$. Vypočítáme tudíž z poslední rovnice

$$\sigma^2(x) = \frac{r(N-1)}{N(r-1)}\,\sigma_{x,v}^2$$

a tento odhad dosadíme nyní do rovnice pro směrodatnou odchylku σ_P výběrových průměrů (13), takže dostaneme

$$\sigma_P = \sigma_{x,v}\sqrt{\frac{N-r}{N(r-1)}}, \tag{45}$$

kde $\sigma_{x,v}$ je směrodatná odchylka pozorovaného náhodného výběru. Vyskytuje se ovšem v této formuli N čili rozsah základního souboru. Výraz (45) dává nejvhodnější odhad směrodatné odchylky rozdělení četností výběrových průměrů, když základní soubor neznáme.

Je-li základní soubor nekonečného rozsahu, redukuje se pro $N \to \infty$ poslední výraz (45) na

$$\sigma_P = \sigma_{x,v} \cdot \sqrt{\frac{1}{r-1}} = \frac{\sigma_{x,v}}{\sqrt{r-1}}. \tag{46}$$

Označíme-li ζ_j odchylku hodnoty pozorovaného znaku od průměru výběrového, je $\sigma_{x,v}^2 = \frac{1}{r}\sum_{j=1}^{r}\zeta_j^2$, takže můžeme rovnici (46) psáti

$$\sigma_P = \sqrt{\frac{\sum_{j=1}^{r}\zeta_j^2}{r-1}} : \sqrt{r}.$$

Čitatel tohoto zlomku

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

se považuje za nejlepší odhad směrodatné odchylky základního souboru, z něhož náhodný výběr o r prvcích byl vzat.

Můžeme tudíž také směrodatnou odchylku výběrových průměrů vyjádřiti vztahem

$$\sigma_P = \frac{\sigma(x, v)}{\sqrt{r}}. \tag{47}$$

Výraz (45) — a výrazy z něho odvozené — je přibližným odhadem, neboť nevíme jak blízko je směrodatná odchylka jednoho pozorovaného výběru $\sigma_{x,v}$, třeba velkého, odmocnině z průměru všech možných rozptylů výběrových (ovšem z téhož základního souboru) $\sqrt{\overline{\sigma^2}}$, za niž jsme ji položili v rovnici (44).

Činíme jen předpoklad, že je jí blízko, což je přijatelnou hypotésou pro výběry velkého rozsahu.

Rovnice (46) a (47) podávají odhad směrodatné odchylky průměru z náhodného výběru rozsahu r, má-li základní soubor nekonečný nebo prakticky velmi velký rozsah. Užívá se jich tedy, když je základní soubor neznámý.

(3,3,l) Příklad. Jeden z náhodných výběrů, uvažovaných v tab. 1 s rozsahem $r = 200$ vykázal průměr $\bar{x}_i = 7{,}03$ a součet čtverců odchylek od průměru $\sum_{j=1}^{r} \zeta_j^2 = 796{,}0$. Jest najíti odhad směrodatné odchylky základního souboru, z něhož byl výběr vzat a odhad směrodatné odchylky výběrových průměrů.

Směrodatná odchylka výběru je

$$\sigma_{x,v} = \sqrt{\frac{796{,}0}{200}} = 1{,}99.$$

Nejvhodnější odhad směrodatné odchylky základního souboru

$$\sigma(x, v) \doteq \sqrt{\frac{796{,}0}{199}} = 2{,}00.$$

Odhad směrodatné odchylky výběrových průměrů

$$\sigma_P = \frac{2{,}00}{\sqrt{200}} = 0{,}14.$$

Jsou tedy meze jedné směrodatné odchylky pro průměr základního souboru

$$\bar{x}_i \pm \sigma_P = 7{,}03 \pm 0{,}14,$$

v nichž je s pravděpodobností 0,683.

(3,4) Metoda největší věrohodnosti. V předcházejícím jsme se pokusili o odhad průměru a směrodatné odchylky základního souboru. Nyní ukážeme jednu z obecných metod, jimiž lze odhadnouti parametry rozdělení četností v základním souboru podle pozorovaného výběru a to takovou, že odhady provedené její pomocí jsou efficientní (za předpokladu, že v daném případě efficientní charakteristika parametru existuje). Ze základního souboru nekonečného rozsahu máme výběr rozsahu r, kde hodnoty znaku pozorovaného na jednotlivých prvcích jsou $x_1, x_2, \ldots, x_j, \ldots, x_r$. Relativní četnost hodnot náhodné proměnné x_i budiž v základním souboru p_i. Jsou-li jednotlivé hodnoty náhodné proměnné vzájemně nezávislé, jak tomu je při rozsahu $N = \infty$, bude pravděpodobnost, že se ve výběru současně vyskytnou hodnoty $x_1, x_2, \ldots, x_j, \ldots, x_r$ právě v tomto pořadí dána podle věty o násobení pravděpodobností součinem $p_1 \cdot p_2 \cdot \ldots \cdot p_r$.

Nezáleží-li pak na pořadí, v němž se vyskytnou jednotlivé hodnoty x_i, nemusíme rozeznávat jednotlivé permutace a pravděpodobnost, že se vyskytne právě takový výběr, jaký máme, bude $P = c \cdot p_1 p_2 \ldots p_r$, kde konstanta c je z kombinatoriky známa.

Předpokládejme, že lze každou relativní četnost v základním souboru p_i vyjádřiti pomocí příslušné hodnoty znaku x_i a parametrů základního souboru čili, že známe tvar rozdělení četností. Pro zjednodušení výkladu tedy předpokládejme, že rozdělení četností v základním souboru je normální, takže se v něm vyskytují jen dva parametry $\bar{x}, \sigma(x)$; potom bude

$$p_i = \frac{1}{\sigma(x)\sqrt{2\pi}}\, e^{-\frac{(x_i - \bar{x})^2}{2\sigma^2(x)}} \qquad (48)$$

a hledaná pravděpodobnost je

$$P = c \frac{1}{(\sigma(x)\sqrt{2\pi})^r} e^{-\frac{1}{2\sigma^2(x)}\{(x_1-\bar{x})^2+\ldots+(x_r-\bar{x})^2\}}. \quad (49)$$

Když hodnoty parametrů neznáme, pak tento výraz pro nějaké předpokládané hodnoty parametrů se nazývá věrohodností (vraisemblance, likelihod) předpokládaných hodnot.

Metoda největší věrohodnosti spočívá pak v tom, že se mají zvolit pro parametry takové hodnoty, pro něž bude věrohodnost maximem, čili hodnoty nejvěrohodnější. Místo maxima výrazu (49) můžeme určit maximum jeho logaritmu, neboť větší číslo má větší logaritmus. Označme předpokládané hodnoty $\bar{x}_v$, $\sigma(x, v)$, potom bude přirozený logaritmus věrohodnosti

$$\lg P = \lg c - \frac{r}{2}\lg 2\pi - r \lg \sigma(x, v) - \frac{1}{2\sigma(x, v)^2}\{(x_1 - \bar{x}_v)^2 + \ldots + (\bar{x}_r - \bar{x}_v)^2\}. \quad (50)$$

Pro určení maxima položíme první derivace podle předpokládaných parametrů rovny nule, tedy

$$\frac{\partial \lg P}{\partial \bar{x}_v} = 0, \qquad \frac{\partial \lg P}{\partial \sigma(x, v)} = 0$$

a dostaneme snadno

$$\bar{x}_v = \frac{1}{r}(x_1 + \ldots + x_r)$$

čili výběrový průměr je odhadem maximální věrohodnosti pro $\bar{x}$. Podobně dostaneme z druhé rovnice, že směrodatná odchylka $\sigma(x, v)$ je nejvěrohodnějším odhadem $\sigma(x)$.

Můžeme nyní formulovati metodu maxima věrohodnosti zcela obecně tím, že místo zvláštní funkce udávající tvar rozdělení četností (48) označíme obecně $p_i = \varphi(x_i, \Theta_1, \Theta_2, \ldots)$, kde Θ_i jsou parametry. Potom bude

$$P = c \prod_{i=1}^{r} \varphi(x_i, \Theta_1, \Theta_2, \ldots),$$

$$\lg P = \lg c + \sum_{i=1}^{r} \{\lg \varphi(x_i, \Theta_1, \Theta_2, \ldots)\}$$

a poněvadž c je konstanta, jedná se o určení maxima funkce

$$L = \sum_{i=1}^{r} \{\lg \varphi(x_i, \Theta_1, \Theta_2, \ldots)\}.$$

Když na př. máme tabulku skupinového rozdělení četností, kde jednotlivé třídní četnosti jsou $n_1, n_2, \ldots, n_l$ a odhad četnosti v j-té třídě pro předpokládaný základní soubor označíme v_j, bude o konstantní člen zkrácený logaritmus věrohodnosti

$$L = \sum_{j=1}^{l} n_j \lg v_j, \tag{51}$$

neboť v tomto případě je

$$P = c \prod_{j=1}^{l} v_j^{n_j}$$

a dále

$$\lg P = \lg c + \sum_{j=1}^{l} n_j \lg v_j.$$

(3,4,l) Příklad. V pozorovaném výběru rozsahu r jsou třídní četnosti $n_0, n_1, \ldots, n_l$. O rozdělení četností v základním souboru učiníme hypotésu, že je vyjádřeno Poissonovou exponentielou, takže absolutní četnosti jsou $\frac{e^{-\lambda}\lambda^x}{x!} \cdot r$ a jejich logaritmy $x \cdot \lg \lambda - \lambda - \lg \frac{x!}{r}$, kde $x = 0, 1, 2, \ldots l$.

Hledáme nejvěrohodnější odhad pro parametr λ a označíme jej λ_v. Podle (51) dostáváme

$$L = \sum_{x=1}^{l} n_x \left\{x \lg \lambda_v - \lambda_v - \lg \frac{x!}{r}\right\}$$

a položíme-li první derivaci podle λ_v rovnu nule, je

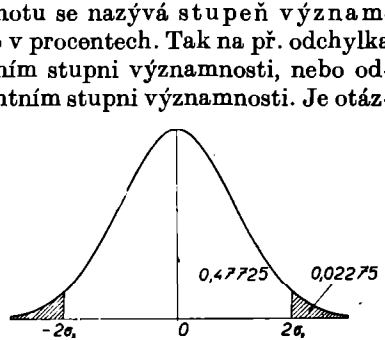
$$\frac{\sum x n_x}{\lambda_v} - \sum n_x = 0 \quad \text{čili} \quad \lambda_v = \frac{\sum x n_x}{\sum n_x}.$$

Je tudíž průměr pozorovaného rozdělení četností efficientním odhadem λ, což je zajímavé potud, že také druhý moment tohoto rozdělení je možným odhadem λ [I. díl, rovnice (57)], takže jsme neměli důvodu považovati jej za horší odhad.

(3,5) Testy významnosti. Viděli jsme již, že užíváme ve statistice metody pracovních hypotés jako v tak mnohých případech vědeckého bádání. Použili jsme na př. předpokladu, že rozdělení četností v základním souboru je normální. Učinili jsme to za tím účelem, abychom převedli zkušenosti získané pozorováním určitého výběru na rozsáhlejší soubor základní, který nemůžeme celý vyšetřiti. Hypotetický soubor s určitými parametry je tu postulátem. Hypotésa je pak přijata, jestliže si vhodným způsobem ověříme, že je možno důvodně soudit, že takový výběr, jaký jsme pozorovali, mohl býti vzat ze základního souboru s těmi vlastnostmi, jež odpovídají hypotése. Není však určité ostré hranice mezi výběry, které by mohly a které by nemohly vyjít ze základního souboru představeného hypotésou, ježto každá výběrová charakteristika má své rozdělení četností, takže vykazuje výběrové odchylky od své očekávané hodnoty, která je parametrem základního souboru. Je možno udati pouze pravděpodobnost, že by z něho mohl vyjít takový výběr, jako je právě pozorovaný. Je-li tato pravděpodobnost malá, hypotésa se zamítne; je-li velká, hypotésa se přijme a odchylka mezi výběrem a hypotetickým základním souborem se připisuje t. zv. náhodnému kolísání výběrovému, které nám na př. pro výběrový průměr znázorňuje tab. 1 ve sloupci 3 a 4.

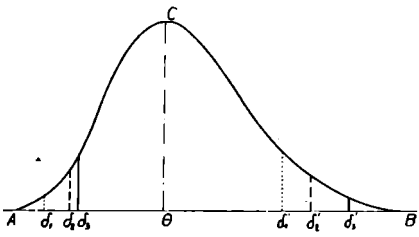
Můžeme-li důvodně předpokládat, že základní soubor má normální rozdělení četností pozorovaného znaku s určitým průměrem a směrodatnou odchylkou, je rozdělení četnosti výběrových průměrů pro určité r normální a můžeme z integrálu Laplace-Gaussova odvoditi pravděpodobnost, že ně-

jaký výběrový průměr se odchýlí od parametru, t. j. od průměru v základním souboru více než o danou veličinu [viz příklad (2,4,1,3)]. Tato pravděpodobnost je na př. 0,0027, že se bude odchylovat nejméně o $\pm 3\sigma_P$, a je tedy malá. Odchyluje-li se tudíž pozorovaný výběrový průměr o trojnásobnou směrodatnou odchylku nebo více, soudíme, že nelze důvodně zařaditi takovou odchylku mezi náhodné a hypotésu zamítneme. Pravděpodobnost, že nějaká náhodná odchylka přesáhne určitou danou hodnotu se nazývá stupeň významnosti a vyjadřuje se často v procentech. Tak na př. odchylka $\pm 3\sigma_P$ je na 0,27 procentním stupni významnosti, nebo odchylka $\pm\sigma_P$ je na 32procentním stupni významnosti. Je otázka, na kterém stupni máme považovati odchylku za statisticky významnou, t. j. tak velkou, že leží na dosti nízkém stupni pravděpodobnosti, aby vedla k zamítnutí hypotésy o základním souboru. Stalo se zvykem vzíti zřetel k decimálnímu systému a voliti pro tuto pravděpodobnost 0,05 (někdy i 0,01), tedy 5procentní stupeň významnosti. Pro normální křivku odpovídá tomuto stupni přibližně odchylka velikosti dvojnásobné směrodatné odchylky, neboť plocha oddělená pořadnicí v tom bodě je 0,0227 celé plochy křivky na straně kladných odchylek a rovněž tolik na záporné straně (obr. 3). Dvojnásobku směrodatné odchylky $2\sigma_x$ odpovídá také přibližně trojnásobek pravděpodobné chyby $0{,}6745\sigma_x$. Uvažujeme stupeň významnosti 0,05 pro kladnou i zápornou odchylku, takže je složen ze dvou hodnot 0,025 na každé z obou stran od parametru, ježto není obyčejně důvodu, abychom soudili, že odchylka nebo rozdíl má míti jen jedno nebo jen druhé znaménko.



Obr. 3. Testování pomocí normální křivky.

Je-li rozdělení četnosti testované charakteristiky nesouměrné, je možno několikerým způsobem stanoviti 5% celé plochy křivky. Tak můžeme odděliti 5% celé plochy stejně velkými odchylkami na obě strany od průměru jak je na obr. 4 $\overline{\Theta\delta_1} = \overline{\Theta\delta'_1}$. Nebo můžeme uvažovati část plochy od pořadnice vztyčené v hodnotě parametru Θ (od něhož odchylky testujeme) napravo zvlášť a nalevo rovněž samostatně a od každé z nich oddělíme 5% plochy, takže pořadnice v δ_2 odděluje 5% plochy $AC\Theta$ a pořadnice v δ'_2 také 5% plochy ΘCB; znamená to tedy také 5% celé plochy. Konečně můžeme najíti na každé z obou stran od hodnoty Θ odchylku δ_3 resp. δ'_3, v níž vztyčená pořadnice odděluje 2,5% celé plochy, takže dohromady je to opět 5% celé plochy. Rozdíly ve výsledcích posledních dvou způsobů nejsou prakticky důležité, ale třetí způsob se zdá výhodnějším.



Obr. 4. Testování pomocí křivky nesymetrické.

(3,5,1) Významnost rozdílu mezi dvěma výběrovými průměry. Máme-li rozhodnouti, zda rozdíl mezi průměry dvou pozorovaných náhodných výběrů můžeme pokládati za náhodný či za statisticky významný, musíme především znáti výběrové rozdělení rozdílů mezi průměry náhodných výběrů z téhož základního souboru.

Vyjdeme pak od hypotésy, že pozorované dva výběry rozsahů r_1 resp. r_2 jsou ze základních souborů s týmž průměrem a počítáme pravděpodobnost pozorovaného rozdílu $d = \bar{x}_1 - \bar{x}_2$. Bude-li tato pravděpodobnost menší než 0,05, zamítneme hypotésu.

Představme si, že vezmeme z jednoho základního souboru veliký počet výběrů rozsahu r_1 a také z druhého základního

souboru veliký počet výběrů rozsahu r_2. Pro každý výběr stanovíme průměr a potom pro každý pár výběrů stanovíme rozdíl mezi jejich dvěma průměry, takže dostaneme tolik rozdílů, kolik je párů výběrů a z nich se vytvoří určité rozdělení četností průměrových rozdílů. Bylo odvozeno obecně a je normální s průměrem rovným nule.

Směrodatná odchylka σ_d tohoto rozdělení četností rozdílů d je větší než směrodatná odchylka výběrových průměrů $\sigma_{P,1}$ resp. $\sigma_{P,2}$, neboť víme [I, rovnice (67′)], že $\sigma_d{}^2 = \sigma_{P,1}{}^2 + \sigma_{P,2}{}^2$, jsou-li oba výběry na sobě nezávislé. Pomocí rozptylů základních souborů $\sigma(x_1)$, $\sigma(x_2)$, jakožto jejich parametrů, dostáváme vzhledem k rovnici (24)

$$\sigma_d{}^2 = \frac{\sigma^2(x_1)}{r_1} + \frac{\sigma^2(x_2)}{r_2}. \tag{52}$$

Pojmeme-li do naší hypotésy, že se oba základní soubory shodují čili, že také $\sigma^2(x_1) = \sigma^2(x_2)$, pak při stejném rozsahu obou výběrů $r_1 = r_2$ bude také $\sigma_{P,1}{}^2 = \sigma_{P,2}{}^2 = \sigma_P{}^2$ čili $\sigma_d = \sqrt{2}\sigma_P$. Kriterium dvojnásobné směrodatné odchylky $2\sigma_d$ pro stupeň významnosti 0,05 vede v tomto případě testování výběrových rozdílů k pracovnímu pravidlu, že statisticky významné jsou rozdíly větší než trojnásobek směrodatné odchylky jednotlivého průměru $3\sigma_P$, neboť $2\sigma_d = 2\sqrt{2}\sigma_P$ a $2\sqrt{2} \doteq 3$. Předpokládali jsme v této úvaze, že známe směrodatnou odchylku $\sigma(x)$ základního souboru. Není-li známa, nahradí se odhadem z výběru, čímž je ovšem dáno omezení pro aplikaci teorie velkých výběrů. Když tedy vycházíme od hypotésy, že pozorované dva výběry jsou z téhož základního souboru, užijeme za odhad vhodné kombinace obou výběrových rozptylů; za takovou můžeme užíti výrazu

$$\sigma_{x,v}{}^2 = \frac{r_1\sigma_{x,1}{}^2 + r_2\sigma_{x,2}{}^2}{r_1 + r_2},$$

který dosadíme do (52) za $\sigma^2(x_1)$ a také za $\sigma^2(x_2)$, které jsou stejné, takže dostaneme

$$\frac{r_1\sigma_{x,1}^2 + r_2\sigma_{x,2}^2}{r_1(r_1+r_2)} + \frac{r_1\sigma_{x,1}^2 + r_2\sigma_{x,2}^2}{r_2(r_1+r_2)} =$$
$$= \frac{(r_1+r_2)\, r_1\sigma_{x,1}^2 + (r_1+r_2)\, r_2\sigma_{x,2}^2}{r_1 r_2(r_1+r_2)}$$

a odhad σ_d^2 tedy bude

$$\sigma_{d,v}^2 = \frac{\sigma_{x,1}^2}{r_2} + \frac{\sigma_{x,2}^2}{r_1}.$$

(3,5,2) Příklad. Dva výběry rozsahu $r_1 = r_2 = 200$ vykazují průměry $\bar{x}_1 = 6{,}68$, $\bar{x}_2 = 7{,}37$ a směrodatné odchylky $\sigma_{x,1} = 2{,}05$, $\sigma_{x,2} = 2{,}02$.

Máme považovati rozdíl mezi těmito průměry $d = 0{,}69$ za statisticky významný?

Odhad směrodatné odchylky bude

$$\sigma_{d,v} = \sqrt{\frac{4{,}20 + 4{,}08}{200}} = 0{,}203.$$

Vidíme tedy, že $d > 2\sigma_{d,v}$ a že tedy rozdíl nepovažujeme za náhodný, neboť jen v 5 případech ze sta dostaneme rozdíly větší než 0,406 (srovnej s tabulkou 1 sloupec 3 a 4).

(3,6) Ověřování hypotés. Objasníme ještě odpověď na otázku jak voliti stupeň statistické významnosti pro ověřování hypotés. Podle předešlého výkladu vidíme, že chceme postupovat podle pravidla: hypotésa je přijatelna, když pozorovaná odchylka (nebo rozdíl) je pod daným stupněm významnosti; je-li nad ním, je zamítnuta. Při tomto postupu můžeme rozhodnout nesprávně

1. tím, že zamítneme správnou hypotésu,
2. tím, že přijmeme nesprávnou hypotésu,

nebo můžeme rozhodnout správně tím,

3. že přijmeme správnou hypotésu, nebo
4. že zamítneme nesprávnou hypotésu.

Víme, že žádná hypotésa nemůže býti dokázána s nějakou konečnou platností, takže naše rozhodování má ráz pokusný a naší snahou je především, abychom v dlouhé řadě případů statistické praxe jen v málo případech zamítli správnou hypotésu, abychom tedy poměr počtu případů nesprávného rozhodnutí 1. k počtu všech rozhodnutí 1. a 3. udělali libovolně malým. Toho můžeme dosáhnouti stanovením vysoké hranice odchylky a tedy vysokého stupně významnosti, jíž odpovídá nízká pravděpodobnost. Navržené pravidlo 5procentního stupně významnosti vede tedy k zamítnutí správné hypotésy průměrně jednou v každých 20 pokusech, pro něž jsme zvolili správnou hypotésu. Kdybychom přijali 1procentní stupeň významnosti, takže pravděpodobnost by byla 0,01, přihodilo by se nám takové nesprávné rozhodnutí průměrně jen jednou v každých 100 pokusech. — Tím bychom sice snížili četnost chyby tohoto druhu, ale zvýšili bychom možnost chyby druhého druhu, neboť jsme tím usnadnili možnost přijmouti hypotésu a tedy také přijmout nesprávnou hypotésu čili učiniti chybu (2). Tak tedy bude volba stupně statistické významnosti a tím stanovení kriteria pro přijetí nebo zamítnutí hypotésy jistým kompromisem mezi nebezpečím dvou druhů chyb. Nelze však stanoviti poměr chyb (2) k celkovému počtu případů, v nichž byla zvolena nějaká nesprávná hypotésa; ten bude záležeti značně na tom, jak je hypotésa blízko pravdě a jak je test přísný. Proto je volba kritické hranice významnosti do veliké míry ovládána velikým množstvím vědomostí z oboru, v němž se šetření provádí a jistou vědeckou tradicí. V důsledku této tradice se na př. doporučují jednoduché hypotésy, t. j. takové, které zahrnují málo konstant před složitými.

Bylo by možno zmenšiti vliv chybných úsudků druhého druhu aniž se dotkneme prvého druhu, kdybychom redukovali směrodatnou odchylku charakteristiky, kterou testujeme. Pro určitý stupeň významnosti totiž vede menší směrodatná odchylka σ k příslušné menší odchylce $\tau\sigma$, která leží právě na naší zvolené hranici významnosti a tedy vede

k menší odchylce, kterou nesprávně odmítneme jako nevýznamnou. Směrodatnou odchylku pak lze redukovati [viz rovnice (24), (39), (40), (41)] zvětšením rozsahu výběru.

Musíme míti také stále na paměti, že označíme-li podle zvoleného pravidla nějakou odchylku za nevýznamnou, znamená to spíše, že její významnost není prokázána. Statistická významnost nedává posudek o velikosti nebo praktické důležitosti nějaké odchylky či rozdílu, neboť ty mohou býti posouzeny jen na základě vědomostí z oboru, do něhož předmět šetření spadá. Je-li nějaký průměr výběrový významně odlišný od hodnoty průměru v hypotetickém souboru, nebo je-li rozdíl mezi dvěma výběrovými průměry významný, nemůže statistická teorie osvědčiti příčinu této odchylky. Příčina může býti v tom, že základní soubor je skutečně různý od hypotetického, nebo výběr není vskutku representativním a náhodným. Může tam býti skutečný rozdíl mezi dvěma základními soubory nebo rozdíl ve výběrové technice.

Testování nesouměrnosti pozorovaného rozdělení četnosti vzhledem k normálnímu lze provésti pomocí míry $\sqrt{\beta_1} = \alpha_{x,3}$, která má v prvním přiblížení směrodatnou odchylku $\sqrt{\frac{6}{r}}$, jak jsme již uvedli, a rozdělení četnosti blízké normálnímu, takže dvojnásobná směrodatná odchylka je prakticky na 5procentní hranici významnosti.

(3,7) Náhodný výběr malého rozsahu. Vyložili jsme vhodné metody k testování statistické významnosti, jichž lze užívati pro výběry velkých rozsahů. Nesmíme jich však užívati, testujeme-li na př. významnost mezi průměry malých výběrů. Proto byla odvozena teorie, která dává přesné testy bez ohledu na rozsah výběru, tedy vhodná pro malé výběry. Spočívá v užití přesných výběrových rozdělení četností místo přibližných, jichž se užilo pro velké výběry. Tím jsme se sice zbavili vlivu výběrového rozsahu, ale ve většině dalších výsledků zůstává omezení, že se aplikují jen na veličiny s normálním rozdělením četností.

Charakteristiky odvozené z velkých výběrů dávají spolehlivé odhady parametrů v základním souboru, kdežto malé výběry poskytují chudé odhady. Každá charakteristika má své rozdělení četností. Soudilo se dlouho klamně, že tvar tohoto rozdělení závisí na rozsahu výběru, ale ve skutečnosti se nejedná o počet všech prvků r, nýbrž o počet nezávislých prvků. Máme-li na př. krabičku zápalek, z nichž každá je nějak označena třeba počtem čárek, které jsme na ni tužkou udělali a máme rozdělit zápalky do pěti skupin, můžeme měnit čtyři skupiny, ale pátá je vždy již určena celkovým počtem zápalek. Máme tedy zde jen čtyři volné cesty, jimiž můžeme provádět libovolné skupiny. Počítáme-li z nějakého výběru s rozsahem 25 prvků průměr a máme-li vzíti z téhož základního souboru druhý náhodný výběr 25 prvků, který by měl týž výběrový průměr jako první, můžeme vzíti libovolně jen 24 prvků tedy $(r - 1)$, neboť poslední je již danou podmínkou určen. Pro tyto případy byl zaveden pojem „stupně volnosti", takže v tomto případě máme na př. 24 stupňů volnosti pro odhad charakteristiky za uvedené podmínky.

Shledali jsme již v odstavci (3,3), že nejlepší odhad rozptylu $\sigma^2(x)$ dostaneme, dělíme-li součet čtverců odchylek od průměru počtem stupňů volnosti $r - 1$ a nikoliv počtem pozorování r.

Počet stupňů volnosti se zde rovná počtu odchylek zmenšenému o počet konstant určených z výběru, jichž bylo použito k pevnému stanovení bodu, od něhož jsou odchylky měřeny, tedy o jednu, ježto se našel jen průměr z výběru.

(3,7,1) Příklad. K osvětlení vlivu, který má užití počtu stupňů volnosti při odhadu rozptylu, bylo vzato náhodně $r = 200$ čísel dvojciferných mezi 10 a 50. Jsou to náhodná čísla z tabulek Tippettových (viz str. 73). Z těchto čísel bylo utvořeno 10 skupin, které budeme nazývati varietami, po $s = 20$ číslech (tab. 2). Můžeme si představit celkovou variaci všech těchto čísel ze dvou složek, jednak ze složky variace uvnitř skupin (variet), jednak ze složky mezi skupinami (varietami).

Tabulka 2.

Běžné číslo	I	II	III	IV	V	VI	VII	VIII	IX	X
1	29	45	14	25	11	47	28	35	18	25
2	45	39	49	32	32	36	24	27	16	39
3	16	29	29	18	46	42	10	18	34	24
4	37	11	31	28	44	36	27	44	18	30
5	50	12	19	20	28	38	11	25	30	24
6	10	37	20	44	40	21	42	33	29	36
7	26	44	49	41	27	41	22	49	35	31
8	19	15	15	10	28	26	30	11	35	10
9	24	50	11	43	27	17	17	17	42	13
10	10	14	22	19	11	50	33	39	50	43
11	10	22	23	10	48	30	44	26	21	27
12	48	20	41	13	21	39	32	29	11	20
13	22	46	40	31	44	21	23	16	45	39
14	13	14	12	45	16	46	25	47	18	30
15	28	21	39	39	36	22	27	10	31	18
16	32	15	43	23	42	34	16	20	26	11
17	10	37	31	11	12	50	20	12	34	46
18	11	26	34	22	48	13	47	42	22	43
19	30	29	49	35	30	46	38	50	24	44
20	37	22	37	49	30	47	12	34	42	24
	507	548	608	558	621	702	528	584	581	577

Oba prameny variace budou vyrovnány, byla-li tato čísla vzata náhodně; nebyly by vyrovnány, kdyby některé skupiny (variety) měly na př. všechna malá čísla a jiné zase všechna velká čísla. Náhodný výběr čísel zajišťuje, že tento případ nenastane. Naše úvaha o variaci znamená, že rozptyly uvnitř skupin, mezi skupinami a rozptyl celkový budou stejné až na odchylky v mezích náhodného výběru. Je-li rozptyl mezi skupinami velmi blízko celkovému rozptylu, takže se mu skoro rovná, musí býti také rozptyl uvnitř skupin skoro přesně roven celkovému rozptylu.

Určíme tedy rozptyl uvnitř každé variety a průměr jejich by měl býti velmi blízko rozptylu pro celý výběr, je-li naše metoda správnou. Sestavíme výpočet rozptylů do následu-

jící tabulky 3. Poněvadž v každé skupině je 20 čísel, bude tam 19 stupňů volnosti pro odhad rozptylu. Dělili jsme tedy v sloupci (7) součet čtverců počtem stupňů volnosti a v sloupci (8) rozsahem souboru.

K sestavení tabulky 3 se doporučuje napsati ještě pomocnou tabulku čtverců čísel tabulky 2, která bude míti jako součty v posledním řádku čísla, jež potřebujeme v 3. sloupci tab. 3. Výpočet sloupce 6 se provádí podle vztahu

$$\sum(x - \bar{x}_i)^2 = \sum x^2 - \frac{1}{s}(\sum x)^2,$$

takže je třeba poznamenati si také pomocný sloupec čtverců čísel druhého sloupce tab. 3, abychom z něho mohli psáti pro $s = 20$ sloupec 5.

Tabulka 3.

Varieta	Σx	Σx^2	$\bar{x}_i$	$\frac{1}{s}(\Sigma x)^2$	$\Sigma(x - \bar{x}_i)^2$	$\Sigma(x - \bar{x}_i)^2 : 19$	$\Sigma(x - \bar{x}_i)^2 : 20$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I	507	16 159	25,35	12852,45	3306,55	174,0289	165,3275
II	548	18 110	27,40	15015,20	3094,80	162,8842	154,7400
III	608	21 602	30,40	18483,20	3118,80	164,1474	155,9400
IV	558	18 620	27,90	15568,20	3051,80	160,6210	152,5900
V	621	22 189	31,05	19282,05	2906,95	152,9974	145,3475
VI	702	27 208	35,10	24640,20	2567,80	135,1474	128,3900
VII	528	16 132	26,40	13939,20	2192,80	115,4105	109,6400
VIII	584	20 306	29,20	17052,80	3253,20	171,2210	162,6600
IX	581	19 043	29,05	16878,05	2164,95	113,9447	108,2475
X	577	19 045	28,85	16646,45	2398,55	126,2395	119,9275
I—X	5814	198 414	Průměr sloupce (7) resp. (8)			147,6642	140,2810

Pro samostatný výpočet celkového rozptylu použijeme jednak 199 stupňů volnosti, jednak celý rozsah výběru $r = 200$.

Výpočet $\sum(x - \bar{x})^2$ pak provedeme opět podle vztahu

$\sum(x - \bar{x})^2 = \sum x^2 - \frac{1}{r}(\sum x)^2$, kde $\sum x^2 = 198\,414$, $\frac{1}{r}(\sum x)^2 =$
$= 33\,802\,596 : 200 = 169\,012{,}98$, takže $\sum(x - \bar{x})^2 = 29401{,}02$
a dále $\frac{1}{199}\sum(x - \bar{x})^2 = 147{,}74$, kdežto $\frac{1}{200}\sum(x - \bar{x})^2 =$
$= 147{,}00$.

Dostáváme tedy čtyři výsledky:

Pomocí stupňů volnosti:

Průměrný rozptyl uvnitř skupin 147,66
Celkový rozptyl výběru 147,74

Pomocí rozsahu výběru:

Průměrný rozptyl uvnitř skupin 140,28
Celkový rozptyl výběru 147,00

a vidíme, že při užití počtu stupňů volnosti dává průměrný rozptyl uvnitř skupin hodnotu rovnou 99,94% celkového rozptylu, kdežto pomocí rozsahu výběru je jen 95,43% celkového rozptylu. V tomto případě je tudíž skutečná hodnota podhodnocena o 4,57%. Osvětlili jsme tak správnost odhadu rozptylu pomocí počtu stupňů volnosti.

(3,8) Významnost průměrů. *t*-test. Při testování statistické významnosti odchylky výběrového průměru od předpokládané hodnoty základního souboru (hypotésy) jsme použili té skutečnosti, že poměr odchylky k její směrodatné odchylce (čili odchylka vyjádřená ve směrodatné odchylce jako jednotce) má normální rozdělené četnosti se směrodatnou odchylkou rovnou jednotce.

Označíme-li odchylku $d = \bar{x}_v - \bar{x}$, je tento poměr $d : \sigma_P$, kde $\sigma_P = \sigma(x) : \sqrt{r}$. Předpokládá se tu tedy, že směrodatná odchylka základního souboru $\sigma(x)$ je známa. Chceme-li však v praxi ověřiti významnost nějakého průměru, je odhad $\sigma(x, v)$ získaný z výběru vše, co známe. Jedná-li se o výběr velkého rozsahu, je tento odhad $\sigma(x, v)$ dostatečně blízký svému parametru $\sigma(x)$ a je možno dřívějšího postupu užíti. Je-li výběr malý, je třeba jisté úpravy vzhledem k nepřes-

nosti, kterou zahrnujeme tím, že užijeme $\sigma(x, v)$ místo $\sigma(x)$. Poměr, jímž testujeme významnost odchylky je nyní $t = d : \frac{\sigma(x, v)}{\sqrt{r}}$. Rozdělení četnosti hodnot t nám umožňuje provésti test přesně bez omezení na velký rozsah výběru. Ale toto rozdělení se rozhodně liší od normálního rozdělení četností hodnot $d : \frac{\sigma(x)}{\sqrt{r}}$, když rozsah výběru a tedy počet stupňů volnosti je malý a je prakticky s ním shodné pro velký rozsah výběru. To poprvé vyzvedl „Student" (1908) a příslušné rozdělení t pro výběry z normálního základního souboru a počet n stupňů volnosti t je dáno funkcí

$$y = y_0 \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

Je patrno, že tedy výběrové rozdělení hodnot t je symetrické vzhledem ku $t = 0$ a závisí jen na n, počtu stupňů volnosti, jimiž byla odhadnuta směrodatná odchylka, takže v tomto případě $n = r - 1$, neboť

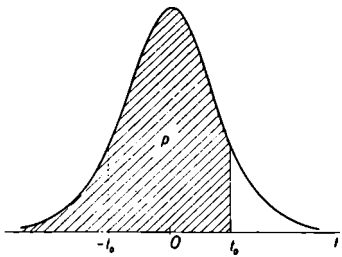
$$\sigma^2(x, v) = \frac{\Sigma(x - \bar{x})^2}{r - 1}.$$

t-křivky odpovídající funkci y mají modus spadající do průměru v $t = 0$, neboť výraz $\left(1 + \frac{t^2}{n}\right)^{-1}$ klesá, když t roste, a na obou stranách jdou větve do nekonečna, jsou jen špičatější ($\beta_3 > 3$) než normální křivka. Také čtenář vidí, že pro $n \to \infty$ spěje výraz $\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$ ku $e^{-\frac{t^2}{2}}$, takže je zřejmo, že pro velká n je t-rozdělení normální. Pravděpodobnost, že při náhodných výběrech dostaneme nějakou hodnotu t, která není větší než t_0, je dána obsahem plochy omezené křivkou, osou t a pořadnicí vztyčenou v bodě t_0; stručně

říkáme plochou křivky až k pořadnici vztyčené v bodě t_0, čili

$$p(t_0) = \int_{-\infty}^{t_0} y_0 \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt$$

obdobně rovnici [I, (50)]. Označíme-li p obecně pravděpodobnost, že pozorovaná hodnota nepřekročí určitou mez t_0, kdežto P bude pravděpodobnost, že pozorovaná hodnota překročí t_0 a to bez ohledu na znaménko, potom bude p plocha křivky nalevo od pořadnice v t_0 (obr. 5), P bude plocha napravo od t_0 plus plocha nalevo od minus t_0 čili (obr. 5) dvojnásobek plochy napravo od t_0, ježto se jedná o křivky symetrické. Bude tedy $P = 2\,(1 - p)$.



Obr. 5. Testování pomocí t-rozdělení.

Sestavíme si několik hodnot pravděpodobností p a P pro srovnání s křivkou normální do tabulky 4.

Tabulka 4.

t	p			$P = 2\,(1 - p)$		
	$n = 10$	$n = 15$	norm.	$n = 10$	$n = 15$	norm.
0	0,500	0,500	0,500	1,000	1,000	1,000
0,6745	0,742	0,745	0,750	0,516	0,510	0,500
1,0	0,830	0,833	0,841	0,340	0,334	0,318
2,0	0,963	0,968	0,977	0,074	0,064	0,046
2,6	0,987	0,990	0,995	0,026	0,020	0,010
3,0	0,993	0,995	0,999	0,014	0,010	0,002

Testujeme-li na 5% stupni významnosti, znamená to, že $P \leqq 0{,}05$, čili $p \geqq 0{,}975$. Jeví se pro praktickou potřebu

testování významnosti účelným sestaviti pro různé stupně volnosti tabulku hodnot t (ležících na některých stupních významnosti na př. 0,1; 0,05 ; 0,01. Vliv větší špičatosti t-křivek proti křivce normální je vážný pro výběry rozsahu menšího než $r = 20$.

Tabulka 5.

Hodnoty t.

P \ n	1	2	3	5	10	15	20	25	∞
0,10	6,31	2,92	2,35	2,02	1,81	1,75	1,73	1,71	1,64
0,05	12,71	4,30	3,18	2,57	2,23	2,13	2,09	2,06	1,96
0,01	63,66	9,93	5,84	4,03	3,17	2,95	2,85	2,79	2,58

Pro velké výběry leží hodnota $t = 2{,}0$ na 5% stupni významnosti: jiné hodnoty t, které jsou na témž stupni, vidíme v tabulce 5 pro $P = 0{,}05$.

Poznámka: Hodnota t je v podstatě poměr nějaké charakteristiky s normálním rozdělením četností a průměrem v nule k odhadu směrodatné odchylky této charakteristiky provedenému s n stupni volnosti. Každý takový poměr, ať vzniká jakkoliv, má totéž výběrové rozdělení četností jako t.

(3,8,I) Příklad. Pevnost nějakého materiálu byla měřena na deseti kusech náhodně vybraných a byly zjištěny hodnoty znaku 63, 63, 66, 67, 68, 69, 70, 70, 71, 71 jednotek. Pomocí dat tohoto náhodného výběru jest ověřiti hypotésu, že průměrná pevnost celého materiálu je 66 jednotek.

Předpokládáme, že rozdělení četností v základním souboru je normální. Výběrový průměr je $\bar{x}_v = 67{,}8$ a odhad směrodatné odchylky $\sigma(x, v) = 3{,}011$. Podle rovnice $t = d : \frac{\sigma(x, v)}{\sqrt{r}}$ bude $t = \frac{67{,}8 - 66}{3{,}011} \sqrt{10} = 1{,}89$. Testujeme-li na 5% stupni významnosti, vidíme z tab. 5, že pro $n = 9$ je hraniční hodnota t větší než 2,23, takže naši pozorovanou hodnotu nepo-

važujeme za statisticky významnou a nemusíme naši hypotésu zamítnouti.

(3,9) Významnost rozdílu mezi průměry. Pomocí t-testu můžeme také ověřovati statistickou významnost rozdílu mezi dvěma výběrovými průměry. Omezíme se na případ, v němž činíme hypotésu, že výběry jsou ze základních souborů, majících společnou směrodatnou odchylku $\sigma(x)$ a týž průměr $\bar{x}$. Poněvadž předpokládáme také normální rozdělení četností základního souboru, znamená tato hypotésa, že výběry jsou z téhož základního souboru. Podle zkušenosti nejsou testy příliš citlivé na mírné odchylky od normálního rozdělení. Při úvaze o těchto testech vycházíme z představy nekonečného základního souboru diferencí, jejichž průměr se rovná nule. Z pozorování jsme dostali dva výběry, jejichž průměry jsou různé, takže tedy vykazují určitou diferenci. Tážeme se, v jakém procentu případů takových dvojic výběrů dostaneme průměrně diferenci tak velkou jako je pozorovaná nebo větší. Testy tohoto druhu musíme rozdělit do dvou skupin.

a) Nejprve máme případ dvou výběrů různého rozsahu r_1 resp. r_2, jejichž prvky jsou na sobě úplně nezávislé, takže hodnoty proměnné netvoří dvojice k sobě nějak vázané. Jsou-li jejich průměry $\bar{x}_1$ resp. $\bar{x}_2$, pak diference $d = \bar{x}_1 - \bar{x}_2$ má normální rozdělení četností kolem nuly a odhad její směrodatné odchylky σ_d provedeme podle rovnice [I, (67′)], která praví, že rozptyl rozdílu dvou proměnných nezávislých se rovná součtu rozptylů každé z nich. Pro odhad rozptylu v základním souboru použijeme kombinace součtu čtverců odchylek od jejich průměrů, kterou dělíme počtem stupňů volnosti $r_1 + r_2 - 2$, neboť dva průměry byly stanoveny, takže bude

$$\sigma^2(x, v) = \frac{\sum_{r_1}(x_1 - \bar{x}_1)^2 + \sum_{r_2}(x_2 - \bar{x}_2)^2}{r_1 + r_2 - 2},$$

směrodatné odchylky průměrů budou (24)

$$\sigma_{P_1} = \frac{\sigma(x, v)}{\sqrt{r_1}}, \quad \sigma_{P_2} = \frac{\sigma(x, v)}{\sqrt{r_2}}$$

a směrodatná odchylka rozdílů průměrů výběrových

$$\sigma_d = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2} = \sqrt{\frac{\sigma^2(x, v)}{r_1} + \frac{\sigma^2(x, v)}{r_2}} =$$

$$= \sigma(x, v) \sqrt{\frac{r_1 + r_2}{r_1 r_2}}. \tag{53}$$

Hodnota t tudíž bude

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma(x, v)} \sqrt{\frac{r_1 r_2}{r_1 + r_2}} \tag{54}$$

a má rozdělení podle křivky t v malých výběrech, ve velkých pak normální rozdělení s jednotkovou směrodatnou odchylkou. Jisté zjednodušení nastává ještě, jsou-li oba výběry stejného rozsahu $r_1 = r_2 = r$, potom

$$\left.\begin{aligned} \sigma^2(x, v) &= \frac{\sum_r (x - \bar{x}_1)^2 + \sum_r (x - \bar{x}_2)^2}{2(r - 1)}, \\ t &= \frac{\bar{x}_1 - \bar{x}_2}{\sigma(x, v)} \sqrt{\frac{r}{2}}. \end{aligned}\right\} \tag{55}$$

Do tabulky 5, hodnot t vstupujeme s počtem stupňů volnosti $n = r_1 + r_2 - 2$.

b) Druhý případ musíme rozeznávat, nejsou-li proměnné nezávislé čili každá hodnota proměnné x_1 je sdružena nějakou logickou cestou s příslušnou hodnotou proměnné x_2 a tvoří tedy dvojice. V takových případech budou míti oba výběry stejný rozsah, takže bude r dvojic hodnot proměnných. Rozptyl nemůžeme stanovit jako v předešlém případě, nýbrž

$$\sigma^2(d, v) = \frac{\sum_r [(x_1 - \bar{x}_1) - (x_2 - \bar{x}_2)]^2}{2(r-1)}$$

a tento výraz můžeme upravit na tvar

$$\frac{1}{2(r-1)} \sum_r [(x_1 - x_2) - (\bar{x}_1 - \bar{x}_2)]^2.$$

Vzhledem k tomu, že součet čtverců lze rozvésti

$$\sum_r (x_1 - x_2)^2 - 2(\bar{x}_1 - \bar{x}_2) \sum_r (x_1 - x_2) + r(\bar{x}_1 - \bar{x}_2)^2 =$$

$$= \sum_r (x_1 - x_2)^2 - 2(\bar{x}_1 - \bar{x}_2) r \left(\frac{\sum_r x_1}{r} - \frac{\sum_r x_2}{r} \right) + r(\bar{x}_1 - \bar{x}_2)^2 =$$

$$= \sum_r (x_1 - x_2)^2 - r(\bar{x}_1 - \bar{x}_2)^2,$$

dostáváme

$$\sigma^2(d, v) = \frac{1}{2(r-1)} \left[\Sigma (x_1 - x_2)^2 - \frac{(\Sigma x_1 - \Sigma x_2)^2}{r} \right]. \quad (56)$$

Hodnota t tedy bude

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma(d, v)} \sqrt{\frac{r}{2}}. \quad (57)$$

(3,9,1) Příklad 1. Ve dvou náhodných výběrech rozsahu $r_1 = 9$, $r_2 = 7$ byly zjištěny průměry $\bar{x}_1 = 196{,}42$, $\bar{x}_2 = 198{,}82$, takže diference $d = 2{,}40$. Byly vzaty nezávisle, takže není důvodu k předpokladu, že jsou v nějaké vázanosti. Součty čtverců odchylek od průměrů jsme stanovili $\Sigma(x_1 - \bar{x}_1)^2 = 26{,}94$, $\Sigma(x_2 - \bar{x}_2)^2 = 18{,}73$, což dává dohromady 45,67. Potom je $\sigma(x, v) = \sqrt{\frac{45{,}67}{14}} = 1{,}81$ a tudíž

$$t = \frac{2{,}40}{1{,}81} \cdot \sqrt{\frac{6}{1}\frac{3}{6}} = 2{,}62.$$

Testujeme-li na 5% stupni významnosti, najdeme v tabulce 5, že pro $n = 14$ je přibližně $t = 2{,}15$, takže rozdíl mezi průměry považujeme za významný.

Příklad 2. Týž druh pšenice vyrostlé ve dvou různých oblastech se zkoumá na obsah proteinu. Z první oblasti bylo pět vzorků s výsledky 12,6; 13,4; 11,9; 12,8; 13,0, z druhé oblasti sedm vzorků s výsledky 13,1; 13,4; 12,8; 13,5; 13,3; 12,7; 12,4. Je tedy v první oblasti průměr $\bar{x}_1 = 12{,}740$ a v druhé $\bar{x}_2 = 13{,}029$. Není-li možno opatřiti další vzorky, jest podle těchto ověřiti, je-li rozdíl mezi průměry významný.

$\Sigma(x_1 - \bar{x}_1)^2 =$	1,2320
$\Sigma(x_2 - \bar{x}_2)^2 =$	0,9943
Celkem	2,2263

$$\sigma(x, v) = \sqrt{\frac{2{,}2263}{10}} = 0{,}472$$

$$t = \frac{0{,}289}{0{,}472}\sqrt{\tfrac{3}{1}\tfrac{5}{2}} = 1{,}047.$$

Z tab. 5 je patrno pro $n = 10$, že tato hodnota $t = 1{,}05$ není významnou na hranici 1% ani 5%, tedy rozdíl mezi průměry $\bar{x}_1 - \bar{x}_2 = -0{,}289$ můžeme pokládati za nevýznamný.

Příklad 3. Byl zkoumán vliv dvou příbuzných krmiv A, B na vývoj párů zvířat a výsledek je obsažen v těchto číslech

Pár	1	2	3	4	5	6	7	8
A	49,2	53,3	50,6	52,0	46,8	50,5	52,1	53,0
B	51,5	54,9	52,2	53,3	51,6	54,1	54,2	53,3

Průměry jsou $\bar{x}_A = 50{,}94$, $\bar{x}_B = 53{,}14$.

Zkoumejme významnost diference $d = 2{,}20$ mezi průměry

a) za předpokladu, že hodnoty pozorované nejsou vázány na dvojice,

b) za předpokladu, že tvoří dvojice.

ad a) Za předpokladu, že výběry jsou na sobě nezávislé, bude

$\Sigma(x - \bar{x}_A)^2 = 32{,}7587$ $\quad \sigma^2(x, v) = 43{,}8974 : 14 = 3{,}1355$

$\Sigma(x - \bar{x}_B)^2 = 11{,}1387$ $\quad \sigma(x, v) = 1{,}77$

Dohromady $43{,}8974$

$$t = \frac{2{,}20}{1{,}77}\sqrt{\tfrac{6}{16}\tfrac{4}{}} = 2{,}486.$$

Z tab. 5 vidíme, interpolujeme-li lineárně pro $n = 14$, že hodnota $t = 2{,}49$ je při testování na hranici 5% významnou, kdežto na hranici 1% není významnou.

ad b) Patří-li stejně očíslované dvojice obou výběrů k sobě, pak dostáváme

$$\Sigma(x_1 - x_2)^2 = 52{,}60, \quad \Sigma x_1 = 407{,}5, \quad \Sigma x_2 = 425{,}1,$$

$$\Sigma x_1 - \Sigma x_2 = -17{,}6, \quad \frac{(\Sigma x_1 - \Sigma x_2)^2}{r} = 38{,}72,$$

$$\sigma^2(d, v) = \frac{52{,}60 - 38{,}72}{14} = 0{,}9914, \quad \sigma(d, v) = 0{,}9957,$$

$$t = \frac{2{,}20}{0{,}9957} \cdot 2 = 4{,}42.$$

V tomto případě je hodnota $t = 4{,}42$ významnou při testování jak na 5% tak na 1% hranici. Považujeme tedy rozdíl $d = 2{,}02$ ve výběrech, v nichž pozorované hodnoty tvoří dvojice za významný.

(3,10) Rozšíření t-testu na tři výběry. Kdybychom měli srovnávati tři výběry nezávislých pozorování, bylo by to jednoduché rozšíření prvního testu. Odhad rozptylu by byl

$$\sigma^2(x, v) = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2 + \Sigma(x_3 - \bar{x}_3)^2}{r_1 + r_2 + r_3 - 3},$$

takže směrodatné odchylky průměrů jsou

$$\frac{\sigma(x, v)}{\sqrt{r_1}}, \quad \frac{\sigma(x, v)}{\sqrt{r_2}}, \quad \frac{\sigma(x, v)}{\sqrt{r_3}}$$

a hodnoty t pro diference mezi průměry budou

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma(x, v)} \sqrt{\frac{r_1 r_2}{r_1 + r_2}}, \qquad t = \frac{\bar{x}_1 - \bar{x}_3}{\sigma(x, v)} \sqrt{\frac{r_1 r_3}{r_1 + r_3}},$$

$$t = \frac{\bar{x}_2 - \bar{x}_3}{\sigma(x, v)} \sqrt{\frac{r_2 r_3}{r_2 + r_3}}.$$

Úloha: Testujte významnost rozdílů mezi průměry variet uvedenými ve 4. sloupci tabulky 3 na 5% hranici významnosti

a) mezi kterýmikoliv dvojicemi,

b) mezi některými trojicemi.

Poznámka: Testujeme-li významnost nějaké průměrové diference, činíme tak na základě určité pravděpodobnosti, že dostaneme tak velké nebo větší diference než je pozorovaná a to buď znaménka kladného nebo záporného. Tak víme, že hodnota t na pětiprocentním stupni významnosti je ta, která odetíná svou pořadnicí 2,5% celkové plochy křivky na pravé straně od parametru a tolikéž na levé straně (obr. 5). Představme si nyní případ, že máme testovati výsledek nějaké setby, při níž bylo semeno nějakým postupem mořeno. Víme již, že tento postup semeni prospívá, takže úrodu zvyšuje; chceme srovnati jeho vliv s výsledkem kontrolní setby, kde tohoto postupu nebylo užito. V takovém případě můžeme uvažovati jen kladné odchylky a pracovati s hranicí významnosti v bodě, kde pořadnice t odetíná 5% plochy křivky jen na kladné straně. Zde se zdá logicky oprávněnějším založiti test na pravděpodobnosti, že dostaneme kladnou diferenci tak velkou nebo větší než je pozorovaná. Podle tabulky hodnot t pak budeme testovati na 5% stupni významnosti, budeme-li bráti ekvivalentní hodnotu t na řádce desetiprocentní, t. j. 0,10.

(3,11) Významnost rozdílu mezi rozptyly. Testujeme-li rozdíly mezi rozptyly výběrů velkých rozsahů, můžeme předpokládati, že rozdíly $\sigma(x_1 v) - \sigma(x_2 v)$ výběrových směrodatných odchylek mají normální rozdělení se směrodatnou odchylkou

$$\sqrt{\frac{\sigma^2(x)}{2r_1} + \frac{\sigma^2(x)}{2r_2}},$$

neboť podle (41) je rozptyl směrodatných odchylek výběrových z normálního základního souboru $\frac{\sigma^2(x)}{2r}$. Neznáme-li rozptyl základního souboru $\sigma^2(x)$, z něhož jsme výběry vzali, nahradíme jej pozorovanými $\sigma^2(x_1 v)$ resp. $\sigma^2(x_2 v)$. Ale v případě výběrů malých rozsahů jsou zase chyby vznikající touto aproximací závažné. Proto se uchylujeme k jinému postupu. Zavádíme nový index

$$z = \tfrac{1}{2} (\lg \sigma^2(x_1, v) - \lg \sigma^2(x_2, v)) = \lg \frac{\sigma(x_1, v)}{\sigma(x_2, v)}, \qquad (59)$$

kde rozptyly $\sigma^2(x_1, v)$ a $\sigma^2(x_2, v)$ jsou počítány pomocí n_1 resp. n_2 stupňů volnosti. Rozdělení četnosti tohoto indexu z bylo odvozeno a má tvar

$$y = y_0 e^{n_1 z} (n_1 e^{2z} + n_2)^{-\frac{1}{2}(n_1+n_2)};$$

obsahuje jako proměnné jen z, n_1, n_2, a je tedy nezávislé na směrodatné odchylce základního souboru $\sigma(x)$. Poněvadž nepotřebuje předpokladů o přibližném vyjádření jejím, lze ho vhodně použít na případy malých výběrů. Abychom si postup osvětlili, všimněme si, že index z se může pohybovat mezi $+\infty$ a $-\infty$, má hodnoty záporné, když $\frac{\sigma(x_1, v)}{\sigma(x_2, v)} < 1$ a kladné pro $\frac{\sigma(x_1, v)}{\sigma(x_2, v)} > 1$. Tvar rozdělení četností je nesymetrický kromě případu $n_1 = n_2$. Kladná část křivky pro $z = \frac{\sigma(x_1, v)}{\sigma(x_2, v)}$ je zřejmě táž jako záporná část křivky pro $z = \frac{\sigma(x_2, v)}{\sigma(x_1, v)}$. Postačí tedy pro jakoukoliv kombinaci stupňů volnosti pravděpodobnostní integrály jen pro kladné odchylky a ostatní dostaneme záměnou n_1 a n_2. Zjednodušíme po-

stup, když se zápornými hodnotami z nepočítáme, ale vezmeme rozdíl logaritmů vždy tak, že je kladný a za n_1 zvolíme ten počet stupňů volnosti, s nímž jsme počítali větší rozptyl. Uvedeme několik hodnot z pro testování na hranici 5%; podrobnější tabulky najde čtenář v [1], str. 150.

Tabulka 6.

n_1 \ n_2	8	10	15	20	30	60	∞
8	0,618	0,561	0,486	0,447	0,409	0,370	0,331
12	0,595	0,535	0,453	0,412	0,369	0,326	0,280
24	0,568	0,504	0,414	0,367	0,318	0,265	0,209
∞	0,537	0,466	0,363	0,306	0,242	0,164	0,000

a hodnoty z na jednoprocentní hranici

Tabulka 7.

n_1 \ n_2	8	10	15	20	30	60	∞
8	0,898	0,810	0,694	0,636	0,577	0,519	0,460
12	0,867	0,774	0,650	0,586	0,522	0,457	0,391
24	0,832	0,732	0,596	0,525	0,452	0,375	0,291
∞	0,790	0,682	0,527	0,442	0,348	0,235	0,000

Pro vylíčený zjednodušený postup si musíme uvědomiti, že tyto hodnoty nejsou příslušnými hranicemi významnosti v našem smyslu. Podle našich úvah leží na 5% stupni významnosti hodnoty z, v nichž vztyčené pořadnice odetínají plochy, z nichž každá je 0,025 celé plochy. Proto leží pětiprocentní hranice významnosti někde mezi pětiprocentními a jednoprocentními hodnotami uvedených tabulek (6 a 7). Je-li počet stupňů volnosti n_1 a n_2 velký, nebo přibližně sobě rovný i když ne velký, blíží se rozdělení z normálnímu se směrodatnou odchylkou $\sqrt{\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ a potom, jak víme,

leží hodnoty z větší než dvojnásobek této směrodatné odchylky nad pětiprocentní hranicí významnosti. Je-li na př. $n_1 = n_2 = 20$, bude směrodatná odchylka 0,224 a z rovnající se dvojnásobku 0,448 je na pětiprocentní hranici čili mezi 0,382 a 0,545, jež dostáváme lineární interpolací z hořejších tabulek.

(3,11,1) Příklad. K objasnění uvedených metod testování při malých výběrech poslouží výsledky studia vlivu insulinu na králíky. Vliv byl měřen procentem svalového glykogenu (hodnota znaku) u 11 zvířat, která byla pod vlivem insulinu a u 10, která nebyla pod tímto vlivem takže slouží za kontrolu. Máme tak měření hodnoty znaku pro dva výběry: a) v kontrole, b) po insulinu.

Hodnota znaku ve výběru	a)	b)
	0,19	0,15
	0,18	0,13
	0,21	0,00
	0,30	0,07
	0,66	0,27
	0,42	0,24
	0,08	0,19
	0,12	0,04
	0,30	0,08
	0,27	0,20
	—	0,12

Průměr ve výběru a) je $\bar{x}_1 = 0{,}273$, ve výběru b) je $\bar{x}_2 = 0{,}135$.

Vzhledem k tomu, že variace znaku od jednoho prvku výběru ke druhému jsou veliké, není rozdíl mezi průměry $d = 0{,}138$ velký. Přezkoumáme proto napřed jeho významnost.

Součet čtverců odchylek

$$\Sigma(x_1 - \bar{x}_1)^2 = 0{,}2530$$
$$\Sigma(x_2 - \bar{x}_2)^2 = 0{,}0715$$
$$\text{Součet} \ldots\ldots 0{,}3245,$$

takže

$$\sigma^2(x, v) = \frac{0{,}3245}{19} = 0{,}01708, \qquad \sigma(x, v) = 0{,}1307.$$

Vzhledem k tomu, že $r_1 = 10$, $r_2 = 11$, bude

$$t = \frac{0{,}138}{0{,}1307}\sqrt{\frac{110}{21}} = 2{,}49.$$

Z tabulky 5 vidíme, že při $n = r_1 + r_2 - 2 = 19$ bude na pětiprocentní hranici významnosti $t = 2{,}09$, takže považujeme naši hodnotu rozdílu za významnou a vliv insulinu na svalový glykogen za skutečný.

Testujeme nyní ještě významnost rozdílu ve variabilitě. Rozptyly jsou $\sigma^2(x_1, v) = 0{,}02811$, $\sigma^2(x_2, v) = 0{,}00715$, takže $z = \frac{1}{2} \lg 3{,}93 = 0{,}684$.

Musíme nyní najíti hodnoty z v tabulkách 6 a 7 pro počet stupňů volnosti $n_1 = 9$, $n_2 = 10$. Poněvadž v nich nejsou hledané hodnoty pro $n_1 = 9$ uvedeny, nýbrž jen pro $n_1 = 8$ a pak až pro $n_1 = 12$, musíme je určit interpolací. Vyjdeme při tom z té skutečnosti, že při témž n_2 jsou změny hodnoty z přibližně úměrné $\frac{1}{n_1}$, t. j. převrácené hodnotě počtu stupňů volnosti n_1.

Tak najdeme napřed z tab. 6 pro $n_2 = 10$ hodnoty z; pro $n_1 = 8$ je v první řádce $z_1 = 0{,}561$ a pro $n'_1 = 12$ na druhé řádce $z'_1 = 0{,}535$, takže rozdíl $\Delta z = 0{,}561 - 0{,}535 = 0{,}026$. Abychom provedli lineární interpolaci vzhledem ku $\frac{1}{n_1}$ najdeme rozdíl $\frac{1}{n_1} - \frac{1}{n'_1} = \Delta = 0{,}1250 - 0{,}0833 = 0{,}0417$ a dostaneme úměru $\left(\frac{1}{n_1} - \frac{1}{9}\right) : \Delta = x : \Delta z$ čili

$$x = \frac{0{,}0139}{0{,}0417} \cdot 0{,}026 = 0{,}009,$$

z čehož plyne, že hledaná hodnota z na pětiprocentní hranici je $z_5 = 0{,}561 - 0{,}009 = 0{,}552$.

Abychom našli z_1, vezmeme z tab. 7 pro $n_2 = 10$ hodnoty z $0{,}810 - 0{,}774 = 0{,}036$ a opět $x = \frac{0{,}0139}{0{,}0417} \cdot 0{,}036 = 0{,}012$ a $z_1 = 0{,}810 - 0{,}012 = 0{,}798$.

Naše hodnota z leží sice mezi pětiprocentní hranicí z_5 a jednoprocentní z_1, ale není jisto, zda je na 5% stupni významnosti, takže můžeme jen říci, že tato data nás vedou k domněnce, že vlivem insulinu mají procenta glykogenu pravidelnější průběh, ale k rozhodnutí by bylo třeba ještě dalších pozorování.

(4) Representativní metoda.

Teorie náhodného výběru tvoří základ t. zv. representativního statistického šetření, jehož hlavním cílem je podati s nejmenším nákladem co nejvíce informací o základním souboru. Representativní šetření je takové, které zkoumá část celého uvažovaného souboru, aby z ní odvodilo úsudky o celku. Částečných šetření se užívá ve velmi různých oborech, kde není možno nebo účelno provésti vyšetření všech prvků souboru, z toho důvodu, že se uspoří peněz a práce nebo se zjednoduší a urychlí šetření i zpracování. Uplatňuje se tak representativní metoda nejen ve vědách přírodních a v technické kontrole výroby průmyslové i zemědělské, nýbrž i v četných oborech hospodářské a sociální stránky života. Tak zjistíme mzdy textilních dělníků nebo kovodělníků representativním šetřením podle určitého výběru a nevyšetřujeme mzdy všech textilních resp. kovodělníků. A jako v otázkách mzdových, tak postupujeme obdobně v otázkách cenových, složení rodin a rozvržení jejich vydání, sledování výroby v různých odvětvích co do množství a jakosti.

V mnohých případech se přeceňoval význam úplného čili vyčerpávajícího statistického šetření, neboť k účelům, pro něž bylo šetření provedeno, stačí hrubší, okrouhlá čísla, uvá-

žíme-li, že přesnost získaných čísel úplným šetřením bývá fikcí.

I kdyby tato čísla byla zcela přesná, jsou takovými jen v okamžiku, k němuž bylo provedeno šetření; čím jsme dále od tohoto okamžiku, tím jsou odchylky od skutečnosti větší. Poněvadž pak ve většině případů v praxi není třeba zcela přesných čísel a na statistických veličinách lpí větší či menší chyby, jak jsme viděli, doporučuje se vždy uvážiti, nemůžeme-li dostati čísla stejné přesnosti nesrovnatelně rychleji a levněji cestou representativní metody čili v obchodní praxi obvyklou a osvědčenou metodou vzorku a má-li v daném případě smysl snažiti se o úplnou přesnost a úplnost materiálu.

Výklad o representativní metodě je vhodno provésti se dvou hledisek. Jednak lze prováděti výběr metodou náhodného výběru, jednak podle principu uváženého čili záměrného výběru.

(4,1) Náhodný výběr s hlediska techniky výběrové. Metoda náhodného výběru může míti dvojí formu: a) neomezeného náhodného výběru, b) oblastního (stratifikovaného) náhodného výběru.

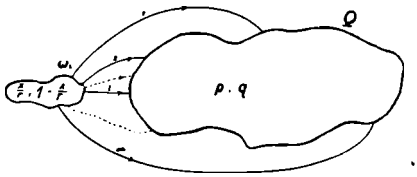
První forma je t. zv. klasická metoda; vybere se z určitého základního souboru jistý počet prvků tak, že pro každý prvek je stejná možnost, aby byl pojat do výběru. Vybere se tedy náhodně jako se konají na př. tahy z osudí. Při tom se postupuje buď tak, že se po tahu vrací prvek zpět, čímž vzniká způsob výběru z nekonečného základního souboru, nebo se vytažený prvek již nevrátí zpět; v tomto případě lze použíti analogie s nekonečným základním souborem jen tehdy, má-li základní soubor veliký rozsah. Jsou-li podmínky náhodnosti a stejné možnosti splněny, není nebezpečí systematické chyby ve výběru.

V příslušné teorii náhodného výběru jsme rozeznávali dva hlavní případy:

a) odhadovali jsme četnost resp. pravděpodobnost p znaku, který se může u prvku vyskytnouti nebo nikoliv (obr. 6),

b) odhadovali jsme parametry rozdělení četností hodnot znaku. Určovali jsme pak pomocí směrodatné odchylky meze odpovídající jistým pravděpodobnostem.

První případ, který se týká alternativního znaku, byl řešen v podstatě rovnicemi (74), (75), (76) I. dílu. Pro druhý případ znaku kvantitativního je dáno řešení v předcházející kapi-



Obr. 6. Parametr a příslušná charakteristika.

tole. V obou případech jsme užili k provedení statistické indukce, metody nazývané nepříliš vhodně empirickou, která nám udává, jak docílíme z hodnoty charakteristiky na př. $f_i = \frac{x}{r}$ nebo $\bar{x}_i, \sigma_i$ pozorované v náhodném výběru ω_i (obr. 6) nejlepší odhad příslušného parametru p, $\bar{x}$, $\sigma(x)$ v základním souboru Ω. Kromě toho jsme podali výklad druhé metody odhadu, t. zv. metody maximální věrohodnosti. Zbývá nám ještě všimnouti si blíže techniky náhodného výběru.

(4,2) Technika náhodného výběru. Výběr representuje základní soubor stručně svými charakteristikami. Jejich hodnoty však podléhají náhodným odchylkám a vedle nich je tu nebezpečí odchylek systematických, které mají své zvláštní příčiny. Provádíme-li na př. statistiku mezd v kovodělném průmyslu pomocí několika závodů místo šetření ve všech závodech, může se státi, že určitá kategorie zaměstnanců má právě v těchto závodech z nějakého důvodu vyšší mzdu než v ostatních. Při zkoumání nákladů na spotřebu domácnosti dělnické či úřednické se užívá domácnostních účtů, jež vede určité množství rodin po dlouhé období; je jasno, že píle a

pečlivost, které je k této práci třeba, vyznačuje již jistou úroveň těchto rodin, takže k takové okolnosti jest přihlížeti při zevšeobecňování úsudků, které vyplynuly z poměrně malého representativního souboru. V náhodném výběru mají býti tyto systematické odchylky vyloučeny. K technickému provedení náhodného výběru se tudíž obyčejně doporučuje nějaký mechanický postup, který by nezávisel na vlivu osoby, jež má výběr sestrojiti. Je-li na př. celý materiál dotazníkový nějakým způsobem seřazen, časově nebo místně, a má se z něho k určitému cíli pořídíti pro rychlé odvození potřebného výsledku representativní výběr náhodně, vezme se podle rozsahu na př. každý dvacátý dotazník, takže rozsah výběru tvoří potom přibližně dvacetinu celého souboru. Jsou-li prvky celého souboru zastupované dotazníky opatřeny pořadovými čísly, je možno opatřiti si miniaturu celého souboru tím, že všechna čísla napíšeme na jednotlivé lístky, jež dáme do osudí a pak z něho vytáhneme náhodně určitý počet lístků; do výběru pak zahrneme prvky souboru, které mají pořadová čísla, jež byla vytažena. Tohoto postupu nelze použíti je-li soubor, z něhož má býti proveden výběr, příliš rozsáhlý, neboť pak přesahuje tento postup často statistikovy praktické možnosti; při rozdělení četností o dvaceti třídách s četnostmi mezi 1 až 60 je potřeba 1000 až 1200 lístků v osudí. Pro takové obsáhlejší případy byla sestavena t. zv. náhodná čísla Tippettova, jichž lze použíti k tvoření náhodných výběrů tak, že se vhodně přiřadí k prvkům základního souboru. Jsou to čísla vzatá ze zpráv o jednom sčítání lidu a číslice jsou kombinovány po čtyřech. Byly pak provedeny pomocí statistických testů zkoušky, které potvrdily, že jsou to čísla vskutku prakticky náhodná. Budtež uvedena ukázkou tato čísla:

2693	1300	5356	7203	1396	1545	9524	4167
5624	3170	5911	7979	9792	3992	6641	2952
7691	6913	1089	3563	2762	3408	7483	2370
8776	4233	8126	6008	6107	1112	5246	0560
6446	8816	6111	7002	9025	1405	9143	2754

Abychom na př. vzali náhodný výběr rozsahu $r = 15$ ze základního souboru tab. 1 rozsahu 1001, očíslujeme jeho prvky číslicemi od 1 do 1001 a najdeme nyní 15 čísel náhodně v mezích od 1 do 1001. Vezmeme tedy některé stránky čísel Tippettových a vybereme na nich prvních 15, která nejsou větší než 1001. Mezi našimi čísly nahoře by to bylo 560 a dále bychom na př. našli 423, 730, 918, 91, 17, 116, 708, 840, 638, 396, 29, 224, 717, 221. Číslování prvků v základním souboru jsme provedli na př. od nejnižší hodnoty znaku nahoru; pak můžeme náhodným prvkům přiřaditi hodnoty znaku, které budou 7, 7, 8, 10, 4, 3, 5, 8, 9, 8, 6, 3, 5, 8, 5, jak vidíme z tab. 1, sloupec (2). Postup je možno také jinak upravit, takže bychom se mohli přiblížit celkovému počtu Tippettových čísel tím, že přiřadíme každému jednotlivému prvku základního souboru 10 čísel. Potom by prvnímu intervalu odpovídala čísla 0000 až 0009, druhému 0010 až 0029 atd. Bližší výklad je uveden v [4]. Také bylo použito k sestrojování náhodných výběrů elektrických strojů třídících.

Máme-li vzíti jako vzorek na př. náhodný výběr r-kusů ze zásilky mnoha beden žárovek, zvolíme především náhodně některé bedny (můžeme pro každou hoditi mincí a bráti jen tu, pro niž padne rub) a v nich zase náhodně různé řady, z nichž žárovku k přezkoušení vezmeme. Kdybychom potřebovali náhodný výběr obyvatelů jedné z hlavních ulic města, můžeme vybrati některá čísla domů, jejichž obyvatelé budou tvořit žádaný výběr. Vyjdeme od některého libovolného domu a vezmeme třeba každý desátý. Nejsou-li tu nějaké zvláštní poměry v pravidelném seskupení zjišťovaných znaků, jako na př. příjem nebo počet členů rodiny, bude zvolená metoda nezávislá na vlastnostech souboru a výběr bude náhodný. Kdyby však v této ulici byl každý desátý dům rohový s velkým obchodem, nebyly by vyšetřované znaky nezávislé na metodě výběru, neboť se obchodní domy vyskytují s touž periodou, jaká byla zvolena pro výběr. Bývá ovšem často obtížno posouditi, zda při provedení výběru nebyl nějaký pramen systematických odchylek, který nemohl býti postřehnut.

Někdy vyžaduje účel šetření, aby byl proveden výběr oblastní. Rozdělí se tedy nejprve celý vyšetřovaný soubor podle určitého znaku na oblasti, a z nich se pak několik prvků vybere náhodně. Tak na př. při shora zmíněném šetření nákladů na spotřebu domácností dělnických na základě domácích účtů je účelno rozděliti všechny dělnické domácnosti podle zaměstnání přednosty domácnosti a sestaviti výběr tak, aby hlavní zaměstnání v něm byla zastoupena způsobem odpovídajícím přibližně struktuře obyvatelstva. V oblasti hlavních druhů zaměstnání vyberou se pak pokud možno náhodně zkoumané domácnosti. Pro vytváření oblastí (strata) může býti směrodatno i více znaků (město, venkov, počet dětí a pod.). Je ještě řada jiných způsobů konstrukce náhodného výběru. Pro vhodnou volbu techniky výběrové v určitém případu musí míti statistik dostatečné znalosti věcné v oboru, do něhož zkoumaný soubor patří a také dosti šťastné intuice.

(4,3) Splnění podmínek náhodného výběru. Praktický význam teorie náhodného výběru je v tom, že umožňuje měřit objektivně chyby odhadu a významnost hodnot zjištěných z náhodného výběru. Můžeme-li pak vzíti několik výběrů z jednoho základního souboru, lze zkoumati, zda rozdělení charakteristik je takové, jak je udává teorie. Odchyluje-li se významně, máme důvod k tomu, abychom zkoumali zvolenou výběrovou techniku a hledali, proč zavádí systematické odchylky. Tento postup předpokládá, že známe rozdělení četností základního souboru, neboť jinak je musíme jen odhadovat podle výběru a pak ovšem nemůžeme toho odhadu užíti ke kritice metody výběrové bez dalšího bližšího vyšetřování. Systematická odchylka od podmínek nutných k sestrojení náhodného výběru musí býti vyloučena dříve než je možno aplikovati výsledky, které plynou z teorie náhodných odchylek výběrových. Aby byly při praktickém provádění sestrojeny podmínky náhodného výběru z pramenných dat, je třeba odvozovati v určitých šetřeních zvláštní schemata k získání náhodného výběru, což závisí na povaze

oboru, v němž se šetření koná. Každý obor vědní nebo druh výroby či obchodu má své problémy při náhodném výběru.

Někde jsou vydána úřední ustanovení pro braní vzorků. Taková jsou na př. vyhlášena výsadní obilní společností o braní vzorků a zkoumání pšenice bohaté na lepek. Vzorky bere osoba úředně k tomu stanovená, která se nejprve přesvědčí, že určité vlastnosti pšenice jsou stejnoměrně vyrovnány v celém množství, z něhož se mají bráti vzorky. Potom se vezme při partiích do deseti pytlů z každého pytle vzorek po 150 g, při partiích do 20 pytlů nejméně 10 vzorků z různých nahodile vzatých pytlů, při partiích do 50 pytlů nejméně 15 vzorků z 15 různých pytlů po 100 g atd. Takto vzaté vzorky se smíchají a stejnoměrně rozprostřou. Celkové množství se potom rozdělí na pět stejných dílů, z nichž se 3 znovu spolu promíchají a tato směs tvoří konečný vzorek, který se rozdělí na tři části nejméně po 250 g, z nichž se každá zapečetí a jedna odevzdá příslušnému zkušebnímu ústavu, druhou dostane prodávající a třetí kupující. Pytle se po odebrání vzorků zaplombují předepsaným způsobem.

(4,4,1) Určení rozsahu výběru při znaku alternativním. Středem praktického provedení representativní metody je určení velikosti výběru čili počtu prvků, které mají býti vzaty z celkového souboru a podrobeny vlastnímu zkoumání. Na této veličině závisí representativní síla výběru a rozhoduje o tom, jak je výsledek dobrý a pravdivý. Viděli jsme, že neusuzujeme z hodnoty relativní četnosti znaku v jednom výběru nebo z průměru jednoho výběru na příslušnou hodnotu v základním souboru, nýbrž vždy jen na jistý obor, v němž může hledaný parametr ležet. Tento obor závisí na stupni pravděpodobnosti, s níž chceme počítat a na počtu prvků zahrnutých do výběru.

Když jsme si předepsali stupeň pravděpodobnosti a rozhodli se pro určité hranice odchylek, pak můžeme rozhodnouti otázku, kolik prvků musí býti pojato do výběru, abychom mohli s určitým stupněm pravděpodobnosti předpokládati, že na př. výběrem zjištěná relativní četnost f bude se

odchylovat od p nahoru či dolů nejvýše o určitou napřed stanovenou veličinu.

Je-li tato napřed stanovená odchylka $z_0 = |f - p|$ a stupeň pravděpodobnosti dán číslem 0,966, vidíme, že odpovídá 1,5násobnému modulu (tab. 8).

Tabulka 8.

t	$\alpha(t)$	$t = j\sqrt{2}$	$\alpha(t)$
1	0,683	$\sqrt{2}$	0,843
2	0,955	$1{,}5\sqrt{2}$	0,966
3	0,997	$2\sqrt{2}$	0,995

Vyjádříme-li odchylku z_0 v jednotce modul, máme známý vztah

$$\gamma = z_0 \sqrt{\frac{r\,\frac{N-1}{N-r}}{2p\,(1-p)}}, \text{ takže } z_0 = \gamma \sqrt{\frac{2p\,(1-p)}{r\,\frac{N-1}{N-r}}}$$

v případě, že se jedná o výběr ze základního souboru konečného a vyňatý prvek se nevrací zpět. Můžeme-li prakticky považovati rozsah základního souboru za nekonečný, je výraz modulu

$$\sqrt{\frac{2p\,(1-p)}{r}}.$$

Je vhodno stanoviti odchylku z_0 jako zlomek hodnoty p, tedy

$$\delta = \frac{z_0}{p}; \text{ je tedy } z_0 = \delta p = \gamma \sqrt{\frac{2p\,(1-p)}{r\,\frac{N-1}{N-r}}}$$

a pro známé hodnoty δ, γ můžeme odtud zjistit jak velké musí býti r, čili jaký zlomek z N prvků musí býti vzat do

výběru. Jednoduchou úpravou dostáváme t. zv. výběrovou rovnici

$$\frac{r}{N} = \frac{1}{1 + \frac{(N-1)\,p\delta^2}{2\gamma^2(1-p)}}. \qquad (60)$$

V našem případě dosadíme za $\gamma = 1{,}5$ a dostaneme jaký zlomek celého souboru nutno vzíti, abychom mohli říci s pravděpodobností 0,966, že relativní četnost bude v mezích $p \pm z_0$ čili v mezích $p \pm \delta p$.

Z výběrové rovnice je zřejmo, že při stálém podílu $\frac{r}{N}$, který se také nazývá výběrovým koeficientem, se δ zmenšuje, jestliže N a p roste. Z toho tedy plyne, že pro znak s větší relativní četností budeme očekávati menší zlomek δ než pro znak s menší relativní četností při stejném poměru rozsahu výběrového r ku N.

Obráceně, má-li býti δ stejné, musíme vzíti pro znak s menší relativní četností větší rozsah výběrový.

Abychom si učinili představu, jakého rozsahu výběrového ze základního souboru rozsahu $N = 100\,000$ je třeba při pravděpodobnosti 0,966 pro různé hodnoty p, aby relativní četnosti f byly v mezích $p \pm \delta p$, sestavíme si několik hodnot do tabulky 9, z níž vidíme na př., že musíme vzíti výběr rozsahu 448 prvků ze základního souboru rozsahu 100 000, abychom mohli očekávati s předpokládanou pravděpodobností 0,966, že relativní četnost nebude kolísati víc než 10% kolem $p = 0{,}5$, že tedy bude mezi 0,45 a 0,55. Když pak dělíme počet všech prvků základního souboru číslem v tabulce, dostaneme kolikátý prvek musí býti vzat do výběru; v našem případě je to tedy $\frac{N}{r} = 100\,000 : 448 = 223$. Hlavní potíž při praktickém užívání výběrové rovnice bývá v tom, že p neznáme. Výběrový koeficient $\frac{r}{N}$ závisí na čtyřech veličinách N, p, δ, γ. Skutečně danou veličinou je N, kdežto pro ostatní tři musíme zvolit určitý předpoklad, abychom

mohli stanovit rozsah výběru. Mezi nimi pak je vnitřní vázanost, neboť p je rozhodující pro velikost modulu a teprve za předpokladu γ-násobku tohoto modulu jako absolutní odchylky může býti určeno δ. Je tudíž zřejmo, že rozsah výběru závisí na předpokladu, který učiníme o velikosti p. Vliv p na velikost výběrového koeficientu vynikne, uvážíme-li jeho extrémní hodnoty. Je-li $p = 1$, čili soubor je jednotný, takže všechny prvky mají uvažovaný znak, je z výběrové rovnice zřejmo, že je třeba výběru krajně nepatrného rozsahu vlastně žádného $\frac{r}{N} \to 0$, ježto každý prvek representuje v tomto smyslu celý soubor. Pro $p = 0$ je $\frac{r}{N} = 1$, čili za předpokladu, že se v celém základním souboru nevyskytuje zkoumaný znak, je třeba velmi rozsáhlého výběru, ba celého základního souboru. Všechny prvky nebo největší část základního souboru musí přijíti do výběru, poněvadž jen tak dojde několik málo prvků se zkoumaným znakem k representaci. Z toho vidíme, že veličina p a r se mění v obráceném poměru čili s poklesem p stoupá výběrovou rovnicí požadovaný rozsah výběru a obráceně. Význam této vlastnosti výběrové rovnice spočívá v tom, že statistik může určiti potřebnou velikost výběru svým předpokladem o veličině p. Z předcházejících vývodů pak vyplývá, že musíme uvažovati tři případy.

Tabulka 9.

p \ 100δ	50	25	10	5
0,01	1.754	6.667	33.333	—
0,10	162	645	4.000	14.286
0,25	54	216	1.333	5.263
0,5	18	72	448	1.785
0,6	12	48	299	1.190
0,7	8	31	193	769
0,8	5	18	112	448
0,9	2	8	50	200

a) Výběr je s hlediska jednoho alternativního znaku representativní svým rozsahem, jestliže použitý předpoklad o p vzatý za základ výpočtu výběrového koeficientu se rovná parametru v základním souboru nebo je menší.

b) Slouží-li výběr k studiu několika alternativních znaků, jejichž parametry jsou p_i, je svým rozsahem representativní, byl-li k určení rozsahu výběru zvolen předpoklad o nejmenším p_i.

c) Při studiu kombinací několika znaků je třeba přihlížeti k minimálním relativním četnostem, které kombinacemi vznikají a jež vedou k velikým rozsahům výběru. V tomto bodě se často hřeší v praxi statistických šetření.

(4,4,2) Určení rozsahu výběru při znaku kvantitativním. Pro znak kvantitativní, nabývající hodnot $x_1, x_2, \ldots, x_l$ obyčejně považujeme výběr za representativní, můžeme-li se spolehnouti s pravděpodobností předem určenou, že odchylky výběrových průměrů od průměru $\bar{x}$ v základním souboru budou v určitých mezích, daných nějakým násobkem směrodatné odchylky nebo modulu. K výběrové rovnici dospějeme, zvolíme-li opět $\delta = \frac{z_0}{\bar{x}}$; poněvadž směrodatná odchylka výběrových průměrů je podle (13) $\sigma_P = \frac{\sigma(x)}{\sqrt{r \frac{N-1}{N-r}}}$, pak má-li se odchylka z_0 rovnat nejvýše γ-násobnému modulu, budeme míti opět vztah

$$\delta \bar{x} = \gamma \sigma(x) \sqrt{2 \frac{N-r}{r(N-1)}},$$

z něhož dostáváme pro výběrový koeficient

$$\frac{r}{N} = \frac{1}{1 + \frac{\bar{x}^2}{\gamma^2 \sigma^2(x)} \frac{N-1}{2} \delta^2}. \tag{61}$$

Můžeme-li považovati prakticky rozsah základního souboru

za nekonečný, je modul vyjádřen výrazem $\frac{\sigma(x)\sqrt{2}}{\sqrt{r}}$, takže pak dostaneme přímo výběrový rozsah

$$r = \frac{2\gamma^2\sigma^2(x)}{\delta^2\bar{x}^2}.$$

Vidíme, že výběrový koeficient závisí na veličinách $\bar{x}$, $\sigma(x)$, γ, δ, N, z nichž první dvě jsou parametry základního souboru, které musíme nějak odhadnouti, což je možno jen podle zkušeností z oboru, jehož se šetření týká, nebo podle dříve provedených analogických šetření.

Poznámka: Dobrým cvičením jest provedení úvah a odvození výběrových rovnic pro odchylky rovnající se nějakému násobku směrodatné odchylky.

Výběrová rovnice nám dala odhad rozsahu pro výběr, který můžeme považovati za representativní po stránce kvantitativní. Representativní síla výběru se však jeví ještě stránkou kvalitativní, což značí, že výběr má představovat pro všechny zkoumané znaky přesné a tedy co nejvěrnější zobrazení základního souboru, který má určitý stupeň homogenity, velkou či malou rozmanitost jednotlivých znaků, která závisí na tom, kolika hodnot může dotyčný znak nabývati a určitý rozsah. Pod těmito třemi hledisky se nám jeví povaha souboru, kterou musíme vždy podrobně uvažovati, abychom zjistili, zda odpovídá povaze souborů, které byly základem odvozené teorie. Soubor s malým rozsahem může poskytovat při úplné homogenitě a malé rozmanitosti přibližně representativní četnost malého počtu znaků ve výběru. Ovšem v rozsáhlém souboru se projeví jednotlivé znaky representativně s jistotou mnohem větší. Rozmanitost přizpůsobujeme tím, že se vzdáme jistých výsledků a zkoumáme jen ty znaky, které vystupují ve velkých četnostech, pro něž tudíž existuje velká pravděpodobnost, že se objeví ve výběru. Do jaké míry je takové přizpůsobení možné, závisí na možnosti omezení po případě

pozměnění účelu šetření. Ve všech případech, v nichž nelze připustiti toto zjednodušení poznatků, musí býti provedeno místo representativního šetření vyčerpávající šetření.

Oblastní výběr vzniká postupem, který rozdělí základní soubor rozsahu N na několik menších souborů rozsahu N_i (obecně různého rozsahu) t. zv. oblastí (strata). Z každé oblasti pak vezmeme technikou náhodného výběru určitý počet prvků r_i zpravidla tak, že poměr $r_i : N_i$ je pro všechna i stejný, což však není nutné. Z těchto částečných výběrů oblastních dostaneme shrnutím celkový výběr. Tento postup se nazývá metodou stratifikovaného čili oblastního výběru. Také můžeme užívati typu oblastního výběru skupinového [1].

(4,5) Záměrný výběr. Podle principu záměrného, také uváženého čili systematického výběru se rozdělí základní soubor na skupiny známých rozsahů a základních parametrů (průměry, směrodatné odchylky). Z těchto skupin se snažíme vybrat záměrně takové, které dohromady poskytnou pro určitou charakteristiku (průměr) týž výsledek jako základní soubor [1]. Matematické podmínky, které jsou podkladem této metody, jsou dosti omezující a teoretická i praktická zkoumání naznačují, že obecně nelze přikládati získaným výsledkům tolik spolehlivosti, jako získaným pomocí náhodného výběru. Výsledky získané representativní metodou se přenášejí na neznámý základní soubor s určitou reservou. Zvláště v oboru hospodářských statistik se vyskytují často representativní šetření, která nejsou vždy provedena průhlednými metodami záměrného výběru a výsledky se bez důkazu vydávají za representativní. Metody náhodného výběru se v tomto oboru neužívá ještě takovou měrou, jak by to dovoloval nynější stav teorie. Větší pronikání lze pozorovati ve statistice výroby zemědělské a průmyslové.

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.

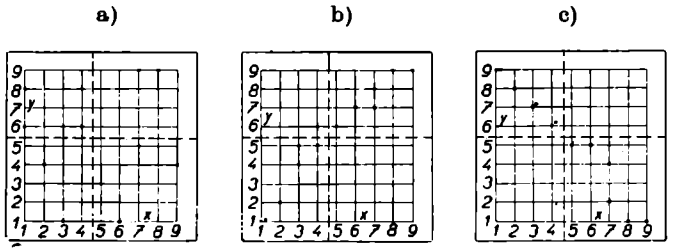
(5,1) Pojem korelace. Představme si, že měříme na prvku určitého souboru hodnoty dvou znaků x a y, takže pro všechny prvky souboru rozsahu r dostaneme r párů hodnot $x_1, y_1; \ldots; x_r, y_r$. Pozorovali jsme na př. páry hodnot uvedené v tab. 10, kde jsou tři různé soubory, každý rozsahu $r = 13$.

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.

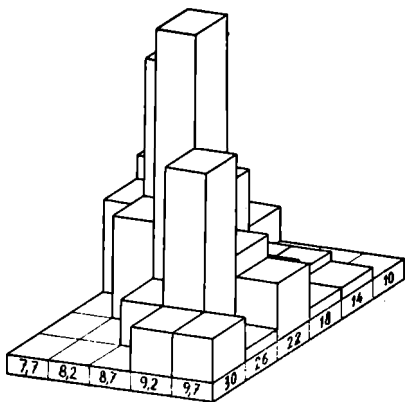
rozsahu, dostáváme roje teček ne nepodobné mléčné dráze. V případech, jakým je náš první soubor, jsou tečky rozhozeny více méně stejnoměrně po celé ploše, takže roj směřuje k tvaru kruhovému, kdežto v případech druhého a třetího souboru se tečky kupí k jedné nebo druhé úhlopříčce.

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

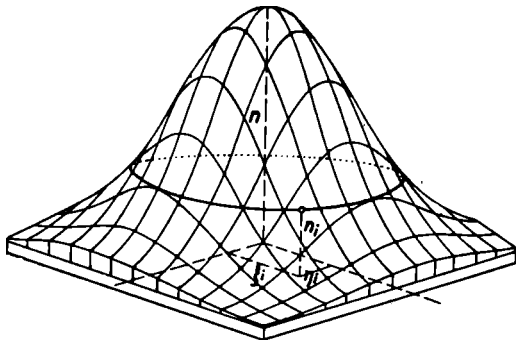
ní plochou dvou proměnných čili normální korelační plochou, která má obdobný význam v teorii dvojrozměrného rozdělení četností jako má normální křivka v teorii rozdělení četností jedné proměnné. Třeba poukázati hlavně na její historický význam, poněvadž s počátku byla teorie korelace budována na předpokladu 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.

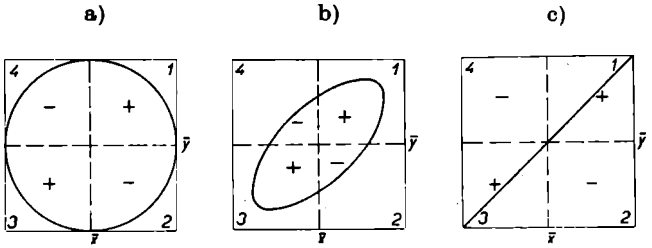
Výsledek tohoto postupu si osvětlíme na našem příkladu. Vypočítáme-li směrodatnou odchylku znaku x, čili délek listů bez ohledu na jejich šířku, použijeme četností krajního (marginálního) sloupce a dostaneme (podle I, str. 22) $\sigma_x = 0,50$. Podobně z marginálního řádku je $\sigma_y = 0,31$. Skutečně 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

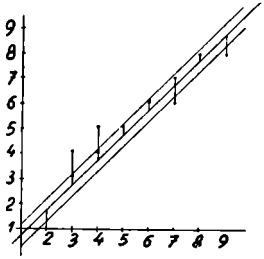
$$y = 1{,}01x + 0{,}72.$$

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

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

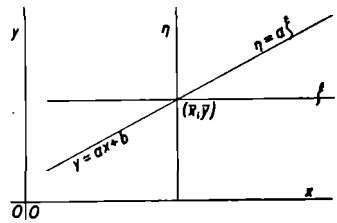
takže budeme psáti

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

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

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

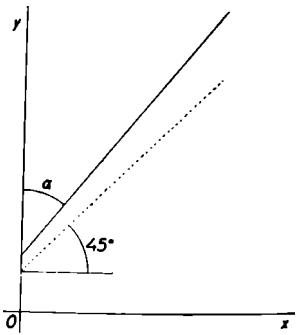
čili $\eta = 5{,}77\xi$. To znamená, že pro každou změnu o 1 v proměnné x nastává změna v y o 5,77.

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

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

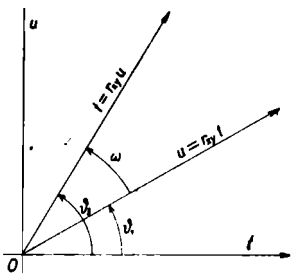
Vyjádříme-li pak rovnice regrese ve směrodatných proměnných $\frac{\xi}{\sigma_x} = t$, $\frac{\eta}{\sigma_y} = u$, dostáváme

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

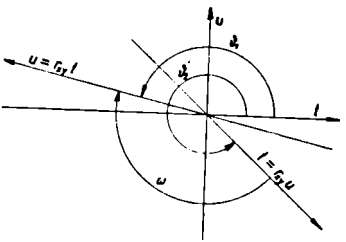
což jsou přímky odhadu vyjádřené v jednotkách směrodatných odchylek a pomocí koeficientu korelace. Jsou to abstraktní proměnné a nezávisí na jednotkách, v nichž byly měřeny původní proměnné x a y. Koeficient korelace je tedy tangentou úhlu, který svírá první přímka regrese s osou t a který označíme ϑ_1 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 $0 \geqq r_{xy} \geqq -1$.

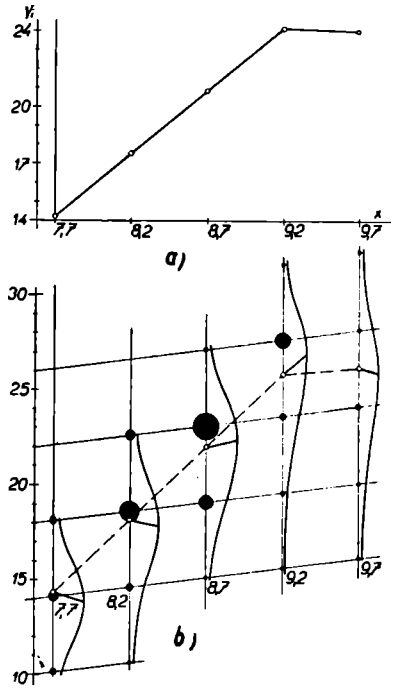
smysl přímek regresních, to jest směr rostoucích hodnot každého znaku. Případ současného růstu obou proměnných je znázorněn v obr. 14, kdežto případ nepřímé korelace je v obr. 15.

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

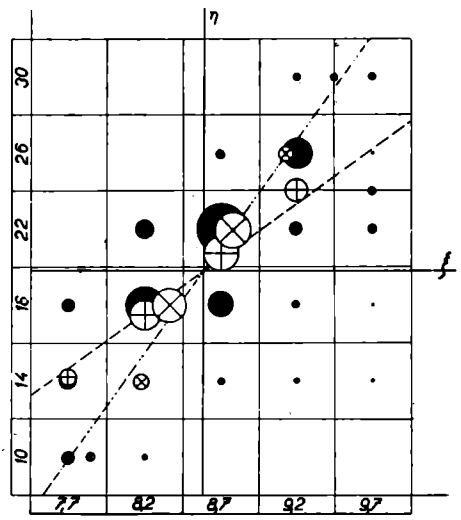
Průměry sloupců

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

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

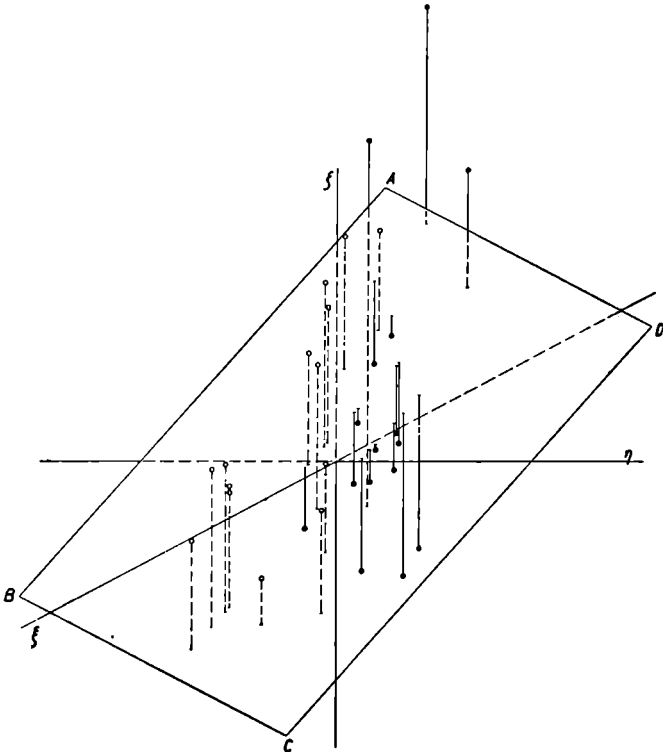
$$s_{zy} = \sigma_z\sqrt{1 - r_{yz}^2} = 3{,}13, \quad s_{zx} = \sigma_z\sqrt{1 - r_{xz}^2} = 3{,}43.$$

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.

ČÁST III.

LITERATURA.

1. *Janko*: Základy statistické indukce, Praha, 1937.
2. *Janko*: Jak vytváří statistika obrazy světa a života. I. díl. Praha, 1942. [Cesta k vědění, sv. 22.]
3. *Rastokin*: Tabulky hodnot $r_{xy} \cdot r_{yx}$ a $\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}$, Praha, 1940.
4. *Florian*: Ukázky ze statistiky experimentální. Rozpravy Jednoty pro vědy pojistné č. 13, Praha, 1934.

OBSAH.

Spisovatel	*Prof. Dr Jaroslav Janko*
Název díla	*JAK VYTVÁŘÍ STATISTIKA OBRAZY SVĚTA A ŽIVOTA. II. díl.*
Vydala	*Jednota českých matematiků a fysiků*
roku	*1944*
Vytiskla	*knihtiskárna Prometheus v Praze*
Komisionář	*nakladatelství Prometheus v Praze*
Vydání	*I.*
Cena	*K 35,—*

vyšší momenty a směrodatné odchylky, všímá si testů významnosti, zmiňuje se o methodě největší věrohodnosti atd. atd. Své výklady doprovází za každým důležitým krokem příklady, aby pojmy a nejdůležitější (základní) statistické pracovní methody čtenáři přiblížil.

Druhá část knihy je věnována statistické *korelaci*, t. j. vzájemnému vztahu několika statistických řad.

Autor tu ozřejmuje, jak se měří korelace, vykládá, co je koeficient korelace, co je regrese a její koeficienty atd., a ukazuje opět na řadě příkladů použití korelace a její hodnocení.

Knížka prof. Janko seznámí vás tak s důležitými pojmy, které byly zavedeny do statistiky v nedávné době jejího rozvoje a umožní vám nahlédnouti do problémů, které statistika zaměstnávají a zároveň vám poskytne všeho toho, čeho je třeba, abyste theorii a methody zde vyložené mohli aplikovati v životě.

K dostání u všech knihkupců

a v nakladatelství

PROMETHEUS.