

F 151275

DR. CYRIL HORÁČEK

# Rukověť

## statistiky

102	1 319	487	489	498	350	522	417	229	74	37	32	6.882
101	703	389	478	548	500	932	822	491	290	132	142	7.676
100	1.640	631	558	526	419	618	455	347	277	189	169	9.279
99	1.217	625	751	786	823	1.139	614	343	210	124	136	8.971
98	682	547	456	394	278	365	196	86	51	27	81	4.706
97	1.077	545	310	236	236	201	66	17	15	11	2	1.835
96	564	68	886	675	506	506	523	11	28	11	8	7.384
95	2.197	568	35	35	35	35	35	35	35	35	35	6.352
94	1.0	8	9	9	9	9	9	9	9	9	9	1.452
93	8	8	8	8	8	8	8	8	8	8	8	8.292
92	1.154	507	561	595	414	414	217	95	60	17	12	6.186
91	1.045	506	116	97	131	93	27	159	49	14	25	7.328
90	284	104	1.012	969	782	981	317	164	90	87	4	4.156
89	745	745	590	473	433	550	307	61	19	8	18	7.797
88	1.075	503	654	640	466	469	185	116	63	24	79	89.124
87	1.105	610	240	325	292	375	208	206	81	49	6	6.097
86	1.289	163	711	731	690	863	435	5.182	2.618	1.297	1.415	13.232
85	378	650	7.602	8.175	7.681	12.393	8.851	5.182	2.618	1.297	1.415	6.097
84	4.642	5.390	483	422	294	364	815	109	29	8	6	7.075
83	3.858	937	1.400	1.537	1.136	769	419	267	181	82	58	8.307
82	479	1.190	480	555	670	529	769	378	176	89	27	2.637
81	966	1.093	680	644	890	808	470	229	77	50	3	7.381
80	917	1.718	927	797	674	719	432	213	169	10	118	13.782
79	947	1.232	675	80	615	716	391	134	40	110	615	23.943
78	764	1.379	927	882	893	1.169	698	409	217	408	7	4.147
77	792	824	927	1.147	896	2.930	1.720	1.069	643	2	1	3.508
76	1.412	1.194	1.779	2.484	2.40	2.22	2.930	1.069	643	2	1	6.195
75	1.478	1.701	1.779	2.484	2.40	2.22	2.930	1.069	643	2	1	23.820
74	415	765	302	37	545	365	135	49	12	5	3	19.460
73	332	618	919	57	545	365	135	49	12	5	3	11.708
72	717	1.056	634	2.137	2.077	1.997	2.608	1.494	724	332	135	6.274
71	2.093	3.277	1.830	1.464	1.708	1.522	2.381	1.536	562	603	429	8.859
70	1.213	1.797	1.145	879	910	785	824	548	567	232	195	8.671
69	702	1.988	689	530	563	495	958	548	567	232	195	7.334
68	250	116	391	781	706	707	819	431	167	71	26	6.391
67	015	1.054	489	1.064	691	686	801	371	154	64	40	4.365
66	000	830	2.784	746	642	497	829	254	150	67	20	5.832
65	553	812	574	690	682	502	360	94	45	35	6	6.040
64	591	878	302	415	444	503	713	996	634	378	176	5.653
63	332	805	467	444	738	854	437	490	240	102	119	7.675
62	643	510	739	415	437	300	4.822	3.244	3.244	3.244	3.244	9.273
61	797	1.343	390	539	539	539	321	332	491	210	124	8.971
60	651	1.129	1.112	1.528	1.934	1.934	321	332	491	210	124	4.706
59	647	839	969	611	544	545	419	618	614	348	51	4.236
58	755	406	880	478	558	526	823	1.139	196	86	17	7.004
57	460	703	282	478	558	526	823	1.139	196	86	17	11.789
56	718	1.640	926	751	786	324	375	201	66	15	35	5.899
55	993	1.093	547	456	310	236	236	201	66	15	35	5.425
54	829	682	345	374	374	374	374	374	374	374	374	6.719



PRAHA

VŠEHRD

R U K O V Ě Ě Š T A T I S T I K Y

*Dr. Cyril Horáček*

# RUKOVĚŤ

# STATISTIKY

VŠEHRD — PRAHA — 1946

## PŘEDMLUVA.

*V této rukověti bylo částečně použito mého spisu Základy statistiky, vydaného v Praze r. 1935, dnes rozebraného.*

*Citace literatury byla omezena na nejmenší nutnou míru, též se zřetелеm k tomu, že nová literatura cizojazyčná není namnoze dosud přístupná.*

*Poněvadž jsem nepovažoval za vhodné použití statistických údajů z doby okupace, a nová statistická data jsou dosud pouze v malé míře k dispozici, byly příklady, pokud se zakládají na domácím materiálu, vybrány z větší části ze statistických dat z doby před rokem 1938, a to opět z různých oborů statistiky sociální.*

*Dojde-li tato rukověť statistiky, obsahující v podstatě pouze teorii statistiky, zájmu též mimo okruh studujících práv, splní svůj účel.*

*Cyril Horáček.*

*V Praze, v červenci 1946.*



## POJEM STATISTIKY.

Jménem statistika se nyní nejčastěji označuje vědecká metoda a theorie, která číselně popisuje a analyzuje hromadné jevy (soubory) a zkoumá vztahy mezi nimi. Slovo statistika má svůj původ ve středověkém latinském slově status, které vedle jiných významů značilo též stát, a od italského slova statista, státník, což souvisí s původním významem statistiky jako nauky o státu (srv. historický přehled). Vytvořen byl tento trochu barbarsky znějící novotvar obdobně jako slova heraldika, numismatika a pod., a časem se vžil ve všech evropských jazycích, třebaže jeho obsah se během doby často pronikavě měnil.

Statistika je charakterisována zvláště těmito znaky:

1. Pracuje pomocí údajů číselných, předmět statistiky musí býti tedy vyjádřen číselně (kvantitativně), jak vyplývá logicky z pojmu hromadný jev.

2. Předmětem statistiky jsou jevy hromadné (kolektivní) neboli soubory. Souborem se nazývá souhrn pozorovaných jevů (skutečností, událostí) stejnorodých, téhož druhu, ale různých vlastností individuelních. Soubory mohou míti nejrůznější povahu: mohou se skládati z pozorování osob, na př. souhrn pozorovaných obyvatelů nějakého státu, žáků středních škol, pachatelů určitého trestního činu, nebo z věcí, na př. souhrn pozorovaných domů, bytů, zemědělských závodů, nebo z jednání a událostí, na př. souhrn pozorovaných trestných činů, nehod, sňatků, úmrtí. Jednotky, z nichž se soubor skládá, musí býti stejného druhu, ale nejsou úplně stejné, neboť se liší vlastnostmi (znaky) individuelních pozorovaných případů, na př. jednotliví obyvatelé věkem, pohlavím, povoláním, národností a jinými vlastnostmi, domy počtem poschodí,

zařazením, sňatky věkem a rodinným stavem snoubenců a pod. (srv. dále kap. 3.).

**Poznámka.** Konkrétní, vymezené soubory, které pozorujeme (někdy se užívá pro soubor též názvu *populace* podle toho, že theorie statistiky zpočátku nejvíce se zabývala soubory obyvatelstva) lze považovati též za pouhé vzorky z universálních pomyslných souborů (*Yule*).

Statistická metoda znamená tedy metodu souborného, kolektivního šetření, čímž se liší zásadně od metody experimentální, kterou se zkoumá jednotlivý jev izolovaný (na př. fyzikální pokus). Statistická metoda je proto nejčastěji používanou metodou pro poznání skutečností sociálních, kde jde převážně o pozorování jevů hromadných, není však na ně omezena, nýbrž lze jí použít na hromadné jevy jakékoli povahy věcné, na př. biologické, antropologické, psychologické, meteorologické atd.

Vzhledem k tomu sluší považovati statistiku za vědu formální nebo metodologické povahy, která je charakterisována jednotným způsobem kolektivního pozorování hromadných jevů (zvláštnost noetické metody). Naproti tomu materiální výsledky statistiky, poznatky získané statistickou metodou na hromadných jevech některého věcně vymezeného oboru, na př. hospodářských, tvoří spolu s ostatními poznatky, získanými jinými metodami noetickými, součást nauky, která se zabývá jevy hospodářskými, t. j. národního hospodářství. Pokusy konstruovati takovou materiální vědu statistickou, která by obsahovala veškeré poznatky, získané statistickým pozorováním hromadných jevů, třeba by se omezovala pouze na hromadné jevy sociální, (ačkoli takovému omezení schází též logické odůvodnění), nutně skončily nezdarem.<sup>1)</sup>

Spor o povahu statistiky související s vývojem statistiky, vyvolal zejména v 19. stol. celou rozsáhlou literaturu, a lze jej nyní považovati za vyřešený ve smyslu výše uvedeném, jakkoli se dosud vyskytují autoři, považující statistiku za vědu povahy věcné (materiální), omezující ji pouze na soubory jevů společenských a definující ji jako vědu o hromadných jevech společenských a o poznatcích, získaných soustavným číselným pozorováním těchto souborů. V české literatuře statistické byl tento směr zastáván Dobroslavem Krejčím.

<sup>1)</sup> O to se pokusil zvláště na počátku 20. stol. německý statistik *Georg v. Mayr*, který se snažil vybudovati celou soustavu poznatků, získaných statistickou metodou v oboru společenských jevů, ve svém díle *Statistik und Gesellschaftslehre* (Tübingen 1914—1922, vyšly však pouze tři díly *Theoretische Statistik*, *Bevölkerungsstatistik*, *Moralstatistik*).

Slovem statistika označuje se však nejen theorie a metoda statistická, nýbrž v obecné mluvě často též výsledky různých statistických šetření, empirická čísla, vzniklá pozorováním nějakého souboru, upravená v tabulkách nebo v jiné formě. Tak se na př. pod označením statistika populační rozumí nejen metoda, kterou pozorujeme soubory populační a nauka o této metodě, nýbrž i číselné údaje o stavu a pohybu obyvatelstva. Někdy se jménem statistika označuje i sama činnost statistická, t. j. provádění nějakého statistického šetření.

## STRUČNÝ PŘEHLED DĚJIN STATISTIKY.

V tomto přehledu vývoje statistiky máme na zřeteli vývoj statistiky jako vědy a nepřehlídíme k úkonům, kterými si opatroval stát určité číselné znalosti k účelům správním, zejména k potřebě vojenské a finanční. Takové soupisy o některých faktech důležitých pro správu státní byly pořizovány již za starověku všude tam, kde některý národ dosáhl vyššího stupně kultury a organisace, jak o tom máme zprávy z Číny, kde již ve třetím tisíciletí před Kristem konaly se soupisy obyvatelstva, podobně ve starém Egyptě a Athénách. V Římě se konaly každých pět let censy, při nichž byl zjišťován počet římských občanů, sloužící v první řadě účelům branným a finančním. Starý zákon obsahuje ve 4. knize Mojžíšově (kap. 1—4) zprávu o sčítání mužů izraelských, způsobilých k boji.

### I. Statistika jako nauka o státu.

Původní význam statistiky byl: všeobecná nauka o státě, zvláště o tom, co je důležité pro řízení státu, což se někdy též označovalo jako nauka o státních znamenitostech. Skutečnosti, které zejména zajímaly státníky, byly forma státní, obyvatelstvo, finance, branná moc, loďstvo atd. Na tento význam statistiky ukazuje též již zmíněná etymologie, statistika znamenala popsání přítomného stavu státu po všech jeho stránkách.

Za spisovatele statistického v tomto významu lze považovati již *Aristotela* (spis *Politeiai*); vlastní zástupci tohoto směru vystoupili na počátku nového věku zvláště v Itálii a potom v Německu, jako t. zv. universitní škola německá.



**Poznámka.** Zmínění italští spisovatelé byli: *Sansovino* (1521—1586, spis *Del governo e amministrazione di diversi regni e repubbliche*) a *Botero* (1540—1617, spis *Le relazioni universali*). Díla tato obsahují popis evropských států a jejich zřízení. Podobná díla vydali ve Francii *Pasquier* (1529—1615, spis *Recherches de la France*), a v Německu *Seckendorf* (1626—1692, spis *Teutscher Fürstenstaat*), a jiní. Z českých spisovatelů bylo by lze sem zařaditi spis *Pavla Stránského ze Záp* (1583—1657) *Res publica Bohemiae* (vyd. 1634).

Jako jednoho z prvních profesorů, který tuto nauku na universitě přednášel, lze uvést *Herrmanna Conringa* v Helmstädtu. Jeho přednášky obsahovaly pod názvem *Notitia rerum publicarum* popis státu po všech jeho stránkách, nauka se blížila obsahem pojmu dnešního politického zeměpisu.

*Conring* sám nenazval sice tyto přednášky statistikou, tohoto označení se však brzy potom počalo užívati všeobecně k označení podobných přednášek, které byly konány na universitách německých v 17. a 18. stol. v souvislosti s učením merkantilistickým, jež se tehdy v Evropě uplatňovalo. Tato disciplína byla zdokonalena zejména *Gottfriedem Achenwallem* (1719—1772), profesorem v Göttingách, který je nazýván též „otcem statistiky“, hlavně proto, že dal nauce o státních znamenitostech jméno statistika, dále poněvadž zavedl do této nauky přesnější systém a získal jí známosti v širších kruzích. Achenwallovo hlavní dílo, rovněž rázu popisného, bylo: *Abriss der neuesten Staatswissenschaft der heutigen vornehmsten europäischen Reiche und Republiken*. *Gottfried Achenwall*

Ze zástupců tohoto směru statistiky, nauky o státech v 18. stol. vynikl zejména *Ludvík Schlözer* (1735—1809), Achenwallův žák a nástupce na universitě v Göttingách. Schlözer pokusil se vymeziti pojem státních znamenitostí, a to v tom smyslu, že nazval tak ony jevy, ve kterých tkvěla síla nebo prameny státní moci, tedy zejména rozlohu státního území, obyvatelstvo, zemědělství, obchod, průmysl, obchodní loďstvo, branné síly atd.

**Poznámka.** Na tomto místě lze ještě uvést jména *Büsching* (1724—1793) a *Crome* (1753—1833), kteří při svých popisech soudobých států používali do jisté míry též číselných údajů (na př. o obyvatelstvu).

## II. Politické aritmetikové.

Jako druhý směr statistiky vedle školy o státních znamenitostech vznikla v Londýně v 17. stol. t. zv. škola politických aritmetiků. Nejvýznamnějšími jejími zástupci byli *John Graunt* a *William Petty*. Jejich práce se zabývaly číselným pozorováním některých jevů sociálních, ze-

jména populačních, byly tedy po stránce methodické bližší nynějšímu pojetí statistiky. *Graunt* (1620—1674), svým povoláním obchodník a samouk, napsal dílo *Natural and political Observations upon the Bills of Mortality*, zabývající se populačními poměry Londýna, již tehdy velkoměsta. Opíraje se o záznamy pohybu obyvatelstva, třeba nedokonalé, pokoušel se odhadnouti počet a vzrůst obyvatelstva londýnského.

Grauntův přítel *Petty* (1623—1687), který byl akademicky vzdělán, byl původcem slova politická aritmetika, jak se nazývá též jeho spis (*Political Arithmetic*), jehož předmět jest rovněž populace londýnská a anglická.

Nedostatek těchto prací se zakládal hlavně na nedostatečném statistickém materiálu, který byl jim podkladem.

Jiným vynikajícím zástupcem této školy byl anglický astronom *Halley* (1656—1742), který na podkladě populačního materiálu města Vratislavě, zpracovaného vratislavským farářem *Kašparem Neumanem*, sestrojil první tabulky úmrtnosti.

V té době počala se tato politická aritmetika zabývatí též jevy hospodářskými. Tu možno uvéstí aspoň jména: *Gregory King* (1648—1712) a *Sir Charles D'Avenant* (1656—1714) v Anglii, a známého francouzského státníka *Vaubana* (1633—1707, spis *Projet d'une dixme royale*).

### III. Další vývoj statistiky až do 19. stol.

Konec 17. století a století 18. znamená též intensivní badání z oboru počtu pravděpodobnosti. Z četných matematiků budiž tu uveden pouze *Jacques Bernouilli* (1654—1705), od něhož pochází t. zv. theorem Bernouilliho (dílo: *Ars coniectandi*). Tento počet pravděpodobnosti byl pak zdokonalen řadou matematiků, zvláště francouzských, z nichž uvádíme jména i odjinud známá: *Laplace*, *Poisson*, *Gauss*, *d'Alembert*.

Významným dílem, které možno nazvati statistickou studií v nynějším slova smyslu, jest spis *Jana Petra Süssmülcha* (1707—1767), pastora v Berlíně: „Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts, aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen“ (1741), v němž zpracoval tehdejší již dosti přesný statistický materiál pruských provincií. Spis obsahuje též pokus o sestrojení úmrtních tabulek pro některé tehdejší státy. Na základě zkoumání úmrtnosti a jejích příčin, poměru pohlaví novorozenců dětí atd., dospěl k poznání některých pravidelností ve vývoji obyvatelstva,

které na něho působily tak mocným dojmem, že viděl v nich řízení božské prozřetelnosti, která tímto způsobem viditelně zasahá do osudů lidstva.

Přes to, že spisy politických aritmetiků a zejména dílo Süßmilchovo ukázalo statistice nové cesty, a přes to, že pokračující zdokonalení státní administrativy poskytlo statistice též nezbytný podklad, statistický materiál, přece vývoj nebyl ještě jednoduchý, neboť proti sobě stály velmi nevraživě obě školy, jednak směru Achenwall-Schlözerova, jednak škola statistiků, kteří používali ve svých studiích údajů číselných, sestavovaných v tabulky (odtud nazýváni byli od německých statistiků Tabellenstatistiker nebo posměšně Tabellenknechte), a spor tento se táhl až do počátku 19. století.

#### IV. Moderní statistika.

Za zakladatele moderní statistiky bývá považován *Adolphe Jacques Quételet* (1796—1874), ředitel observatoře v Bruselu, první president centrální statistické komise belgické. Quételet napsal velkou řadu statistických prací, z nichž budtež jmenovány jen některé význačnější, a zároveň to, co nového Quételet do statistiky přinesl. Ve své knize *Recherches statistiques sur le Royaume des Pays-Bas* (1828) přidržuje se ještě v zásadě metody školy Achenwallovy, nový jest však jeho požadavek, aby statistika snažila se vysvětliti jevy, které pozoruje, tedy, aby podle vzoru věd přírodních hledala příčinné vztahy mezi jevy sociálními. Vycházeje svým školením ze věd přírodních (byl původně astronomem), hledal v jevech společenských podobné zákony, jako zákony přírodní řídí přírodní jevy. Tyto snahy se objevují zvláště v největším díle Quételetově „*Sur l'homme et le développement de ses facultés ou Essai de physique sociale*“.

Quételet pevně určil předmět statistiky, hromadné jevy sociální, převzav metody k jejich zkoumání v podstatě od politických aritmetiků. Odtud nazývá se též otcem moderní statistiky. Pravidelnosti, které se objevují při statistickém zkoumání hromadných jevů sociálních, Quételet přeceňoval hlavně v důsledku toho, že neměl k dispozici statistický materiál dosti dlouhodobý; další práce rázu spíše antropometrického vedly jej ke konstrukci „průměrného člověka“ (*homme moyen*), který měl vyjadřovati jakousi personifikaci průměru (střední hodnoty) fyzických i duševních vlastností příslušníků svého národa, a ke snaze o řešení problému svobodné vůle člověka pomocí statistiky ve smyslu deterministickém.



Po Quételetovi, od druhé poloviny 19. stol., vyvíjela se statistika jako nauka o hromadných jevech. Úplné jednoty o pojmu statistiky nebylo však dosaženo, a dosud se vyskytují ve statistické theorii směry odlišné, různící se zejména v tom, zda zdůrazňují více formální nebo materiální stránku statistiky. Krajiním směrem je t. zv. matematická statistika, která považuje statistiku za zvláštní případ použité (aplikované) matematiky, nedbajíc toho, že statistika není naukou abstraktní, nýbrž empirickou.

Zdokonalením a zjemněním statistických method se velmi rozšířilo jejich používání v různých oborech vědních. K tomu přistoupilo ještě zvýšené použití statistické metody na jevy společenské, zejména hospodářské, pro které je statistická metoda nepostradatelná. Kdežto druhá polovice 19. stol. s převládajícím liberalismem politickým i hospodářským nebyla příliš příznivá statistickým šetřením a v důsledku toho ani statistickému badání, znamenají poslední tři desetiletí podstatnou změnu názorů na úkoly státu na poli sociálním a hospodářském ve směru dalekosáhlých zásahů státu do života občanů ve všech oborech. Ve všech státech byla uskutečněna ve větší či menší míře rozsáhlá opatření sociálně-politická i pronikavé zásahy státu do výroby, distribuce i spotřeby, takže přítomné období národního hospodářství bývá charakterisováno jako hospodářství řízené. Všechny tyto úkoly státu vyžadují, aby poměry populační, hospodářské a sociální byly objasněny co nejvíce a co nejpodrobněji jako předpoklady pro státní politiku (srv. hospodářské plánování) a stejně i účinky zásahů státu do těchto poměrů. Přitom připadá hlavní úkol statistice a odtud její zvýšený význam, ovšem též zvýšené nebezpečí jejího zneužití.

**Literatura ke kapitole 1—2:** České systémy a učebnice statistiky: *J. Beneš* O statistice (1920), *D. Krejčí* Základy statistiky zvláště pro zemědělce a družstevníky (2. vyd. 1923), *St. Kohn* Základy theorie statistické metody (1929), *O. Horáček* Základy statistiky (1935), *J. Janko* Jak vytváří statistika obrazy světa a života (I. díl 1942, II. díl 1944). Dále učebnice význačného anglického statistika *Yule-a* Úvod do theorie statistiky v českém překladu *Dr. V. Nováka a J. Mráze* (1926). K dějinám statistiky: *Westergaard* Contributions to the history of Statistics, Londýn 1932.



## LOGICKÝ PODKLAD STATISTIKY.

### I. Pojem statistického souboru.

Pojem jevů stejnorodých, stejného druhu, na němž se zakládá statistický soubor, který je vlastním předmětem statistiky, je pojmem pouze relativním. Zakládá se na abstrakci od rozdílných vlastností (znaků) individuálních, obsažených v jednotkách, z nichž se soubor skládá. Širší nebo užší měrou abstrakce, které použijeme, rozšiřuje se nebo zužuje též statistický soubor, ovšem čím je širší mez abstrakce, tím se stává stejnorodost (homogenita) souboru pochybnější. Povahou statistiky je dáno, že netvoří vždy samostatných definic pro pojmy jevů, které jsou předmětem statistického šetření, nýbrž že obvykle přejímá pojmy už hotové, vytvořené jinými vědami, na př. pojmy Čech, zemědělec, vystěhovalec, výroba, spotřeba, důchod, průmyslový podnik a pod. V důsledku toho nelze tyto pojmy měnit a soubor rozšiřovati nebo zužovati libovolně, nýbrž toliko vybíráním a připojováním dalších znaků, které lze statisticky zachytiti, na př. k pojmu vystěhovalec můžeme připojiti znaky: mužského pohlaví, české národnosti, římsko-katolického vyznání, povolání zemědělského atd. Tímto omezováním (zužováním) tvoříme z celkového souboru soubory dílčí. Čím více tímto způsobem soubor omezujeme, tím se stává stejnorodějším, avšak tím se zároveň zmenšuje jeho rozsah. Z povahy hromadného jevu však vyplývá, že se soubor musí skládati z velkého počtu případů (jednotek), jak to pak vyjadřuje pravidlo zvané zákon velkého čísla.

Soubor musí býti vymezen *věcně* dříve než se počne konati statistické šetření, srv. kap. 5.

Pokud jde o časové vymezení souborů, rozeznáváme dvojí druh souborů: Soubory jevů trvalých, v čase rozložených (ovšem trvání jejich je pouze omezené a tedy též jen relativní, neboť jednotky souboru časem zanikají a jiné do něho vstupují). Soubory tyto lze statisticky zachytiti v jedinou dobu, theoreticky v jediný okamžik, na př. soubor obyvatelstva, soubor závodů zemědělských a pod. (dobře věc vystihuje německý termín Bestandmassen). Jiný druh souborů jsou soubory jevů okamžitých, událostí, jejichž trvání je omezeno na okamžik, a které jsou statisticky zachycovány nepřetržitě během určitého časového období, na př. soubor úmrtí zachycován je nepřetržitě během jednoho roku (německý termín je Ereignismassen).

Soubory musí býti dále vymezeny místně, na př. státním územím, na němž je rozložen soubor obyvatelstva, obvodem města, v němž je rozložen soubor domů; není-li místně vymezen jiným způsobem, bývá určen místním obvodem kompetence orgánu, který statistické šetření podniká. Podle povahy souborů tvoří se pak případně přirozená vymezení místních souborů, na př. přirozené oblasti rovinaté nebo hornaté, oblasti přímořské, bezlesé, stepní, nebo vymezení hospodářská, na př. oblasti zemědělské se stejnými podmínkami pro výrobu zemědělskou, městské aglomerace a pod.

**Poznámka.** Zeela konkrétně můžeme si představit statistický soubor jevů trvalých: ulehne-li za letního dne do stínu košaté vzrostlé lípy, vidíme v její koruně statisíce listů, které všechny jsou téhož druhu, téže podoby, charakteristického srdčitého tvaru. Přihlédneme-li však blíže k jednotlivým listům, sotva bychom našli dva listy úplně se shodující co do délky, šířky, rozměrů vroubkování, použijeme-li jemných měřítel srovnávacích. Listy téhož stromu jsou všechny jednoho druhu, ale různých vlastností individuálních.

Ovšem můžeme též považovati tuto lípu za náhodně vybraný vzorek ze všech lip toho druhu rostoucích po zeměkouli, t. j. myšleného universálního souboru všech existujících lipových listů.

## II. Poměr statistické metody k indukci.

Jestliže můžeme metody vedoucí k našemu poznání, metody noetické, rozdělit na dvě hlavní skupiny:

- (1) Metodu deduktivní neboli spekulativní, která se snaží abstraktními logickými pochody dospěti k poznání;
- (2) Metodu induktivní v nejširším slova smyslu neboli empirickou, která na základě zkoumání skutečností se snaží dospěti k všeobecnému poznání, od poznání jednotlivostí dochází se k úsudku o celku, potom

statistickou metodu zařadíme do skupiny druhé, poněvadž také při ní vycházíme od pozorování jednotlivých jevů a od tohoto pozorování docházíme k úsudku o celku, statistickém souboru.

Statistickou metodu je však považovati za zvláštní druh úsudku indukčního. Schematicky mohli bychom si vše představit takto:

Pozorujeme-li opětovně, že určité podmínky, které označíme písmeny  $A, B, C$ , mnohokráte vedly k témuž důsledku, označenému písmenem  $D$ , mluvíme o úsudku indukčním, podle něhož usuzujeme, že kdykoli se vyskytne  $A, B, C$ , vždy bude následovati  $D$ . Druh těchto vztahů snažíme se pak dále zkoumati v tom směru, zda jde o příčinnou závislost (kausalitu). Jestliže máme dva členy  $S$  a  $T$  spojené v úsudku o téže jevu, vyjadřujeme při indukčním úsudku podmíněnost obou členů ve formě: jestliže je  $S$ , je i  $T$ .

K rozhodnutí otázky, který z obou jevů  $S$  a  $T$  je podmínkou a který důsledkem, slouží dvě kriteria: výskyt jevu  $S$  je podmínkou jevu  $T$ , jestliže výskyt jevu  $S$  pravidelně s sebou přináší výskyt jevu  $T$ , avšak nikoli naopak, a jestliže v časovém sledu výskyt jevu  $S$  předchází výskytu jevu  $T$ . Zvláštním případem je vzájemná příčinnost obou jevů, jestliže i jev  $S$  podmiňuje jev  $T$ , i jev  $T$  podmiňuje jev  $S$ .

Tento úsudek však předpokládá, že známe veškeré podmínky a veškeré důsledky, které jim následují. Pouze v tomto případě lze užití obecné metody indukční, umožňující nám objasniti příčinný vztah. Naproti tomu všude tam, kde nejsou známy buď veškeré podmínky nebo veškeré důsledky nebo obojí (tak tomu bude pravidelně u jevů sociálních), nutno použít metody statistické, která se zakládá na logickém postupu poněkud odlišném.

Schematicky můžeme si v takovém případě označiti podmínky písmeny  $A, B, C + X$ , kde  $X$  znamená jednu nebo i více podmínek neznámých, výsledný jev pak písmenem  $D + Y$ , kde  $Y$  značí opět jeden nebo více důsledků neznámých.

Kdežto metoda indukční umožňuje založiti konkluse na jediném pozorování nebo experimentu, který ovšem se může libovolně opakovati, a může poskytnouti závěr o příčinné souvislosti, statistická metoda se zakládá na hromadném pozorování a poskytuje konkluse toliko o pravděpodobných souvislostech. Úsudek zakládající se na statistické metodě je tedy považovati buď za zvláštní druh úsudku indukčního (někdy se též mluví o neúplné indukci) nebo za samostatný postup noetický, zakládající se na zvláštním pochodu logickém.



**Příklad.** Methody indukce lze užíti na objasnění nějakého jevu fyzikálního, na př. volného pádu hmotného tělesa, jenž nastává vždy, uvolníme-li je a jenž se řídí pravidlem, že dráha, kterou těleso urazí, je úměrná dvojmuči času, po který těleso padá. Naproti tomu nelze předem určití pohlaví novorozeného dítěte v konkrétním případě, nýbrž toliko pravděpodobnost rozložení pohlaví mnoha novorozených dětí na základě předcházejícího pozorování methodou statistickou.

Mohli bychom se postavit na stanovisko: Kdybychom nebyli nevědomí, nebylo by pravděpodobnosti, nýbrž toliko jistota. Ovšem ani naše nevědomost není absolutní, my něco o jevu víme, jinak by nebylo ani pravděpodobnosti. Pojem pravděpodobnosti souvisí takto s naší nedostatečnou znalostí věcí (Henri Poincaré).

### III. Pojem pravděpodobnosti.

Logický pojem pravděpodobnosti je v úzké souvislosti s formou hypoteticko-disjunktivního úsudku. Schema hypoteticko-disjunktivního úsudku zní: Na praemissu  $A$  může následovati konkluse  $B$  nebo  $C$  nebo  $D$  nebo  $X$ . Tedy kdykoli nastane  $A$ , nastane s určitostí jedna z těchto konklusí, praemissa  $A$  jest k úhrnu těchto konklusí v poměru nutnosti, avšak k jednotlivým konklusím (totiž  $B$  nebo  $C$  nebo  $D$  nebo  $X$ ) pouze v poměru *možnosti*. Úkolem počtu pravděpodobnosti jest pak vyšetřiti stupeň této možnosti.

Chceme-li tedy zjistiti stupeň možnosti, že nastane jev  $B$ , předchází-li jev  $A$ , a je-li znám celkový počet všech možných, ale vzájemně se vylučujících jevů, které mohou následovati, a ona část možností, která jediná vede k očekávanému jevu  $B$  (jinak vyjádřeno: je příznivá jevu  $B$ ), pak poměr obou těchto veličin dává nám vzorec matematické pravděpodobnosti, že nastane  $B$ . Tento poměr se vyjadřuje zlomkem: 
$$\frac{\text{počet případů příznivých}}{\text{počet případů možných}}$$
 a nazývá se pravděpodobností. V theorii statistiky se označuje matematická pravděpodobnost obvykle písmenem  $p$  (probabilitas), jehož bude nadále užíváno.

Krajní mez této pravděpodobnosti je jednak případ, kdy nevyskytuje se žádný případ příznivý, pak i pravděpodobnost  $p = 0$ , mluvíme o nemožnosti toho, že by se případ  $B$  mohl vyskytnouti. Jestliže naopak všechny případy možné jsou zároveň případy příznivými, pak pravděpodobnost  $p = 1$ , nastává druhá mez, mluvíme o jistotě, že případ  $B$  se vyskytne. Konečně lze vytknouti ještě případ střední, kdy počet



případů pro jev  $B$  příznivých je stejný jako počet případů pro jev  $B$  nepříznivých, pravděpodobnost  $p = \frac{1}{2}$ , což se někdy nazývá pravděpodobností v užším smyslu.

Příklady takové pravděpodobnosti poskytují zvláště t. zv. hry náhodné, jako jsou kostky, loterie, ruleta, vrh penízem a pod., t. j. takové hry, kde činnost jako hod, tah a pod. může vésti k několika výsledkům, jejichž pravděpodobnost je předem stanovena. Na př. u kostek pravděpodobnost, že padne určitá stěna s určitým číslem, rovná se  $p = \frac{1}{6}$ , poněvadž je šest stěn a tedy šest případů možných, u vrhu mincí pravděpodobnost, že padne líc nebo rub  $p = \frac{1}{2}$  atd.

**Poznámka.** Pokud jde o výklad a odvození několika základních pravidel počtu pravděpodobností, pravidla o sčítání pravděpodobností, očekáváme-li alternativní výsledek, pravidla o násobení pravděpodobností, očekáváme-li současně vyskytnutí několika jevů, pravidla o násobení pravděpodobností při jevech na sobě závislých, poukazuje se na učebnice algebry používané na vyšších středních školách a příklady v nich uvedené.

$$\text{pravděpodobnost} = \frac{\text{počet příp. příjvých}}{\text{počet příp. možných}} = p (\text{probability})$$

## THEORETICKÉ ROZLOŽENÍ SOUBORU A ZÁKON VELKÉHO ČÍSLA.

### I. Binomické rozložení.

Vedle všeobecné pravděpodobnosti, že určitý jev nastane za určitých předpokladů, jde ještě o to, určití *pravděpodobnost toho, že určitý jev se bude opakovati v určitém počtu při daném množství pozorování.*

Označíme-li pravděpodobnost toho, že jev  $A$  nastane písmenem  $p$ , pravděpodobnost, že nastane jev  $B$ , písmenem  $q$  (při čemž  $A$  a  $B$  se vzájemně vylučují a dohromady vyčerpávají všechny možné případy, tedy  $p + q = 1$ ), počet pozorování označíme písmenem  $n$ , a konečně písmenem  $m$  počet výskytů  $A$ , který očekáváme při tomto počtu pozorování, potom pravděpodobnost, že jev  $A$  se  $m$ krát vyskytne při  $n$  pozorování, je dána vzorcem:

$$P = \binom{n}{m} p^m q^{n-m},$$

při čemž symbol  $\binom{n}{m}$  značí zlomek  $\frac{n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \dots \cdot (n-m+1)}{1 \times 2 \times 3 \times 4 \times \dots \times m}$

**Příklad č. 1.** Je dána urna se stejným počtem koulí červených a černých (třeba 5 a 5). Pravděpodobnost, že se při jednom tahu vytáhne koule červená, rovná se tedy  $p = \frac{1}{2}$ . Pravděpodobnost, že nastane druhý jedině možný jev, to je, že vytáhneme kouli černou, rovná se rovněž  $q = \frac{1}{2}$ . Pravděpodobnost pak, že při deseti tazích vytáhneme pět koulí

červených, dána je vzorcem výše uvedeným, kde dosadíme:

$$n = 10, m = 5,$$

$$\text{potom } P = \binom{10}{5} \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^5 = \frac{10 \times 9 \times 8 \times 7 \times 6}{1 \times 2 \times 3 \times 4 \times 5} \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^5 = 0,246$$

Za stejných předpokladů je pravděpodobnost, že při deseti tazích vytáhneme osm koulí červených:

$$P = \binom{10}{8} \times \left(\frac{1}{2}\right)^8 \times \left(\frac{1}{2}\right)^2 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3}{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8} \times \left(\frac{1}{2}\right)^8 \times \left(\frac{1}{2}\right)^2 = \frac{45}{1024} = 0,0439.$$

Sestavíme-li pro týž příklad tabulku, ve které jsou vyznačeny veškeré pravděpodobnosti pro všechny případy opakování očekávané události, tedy pravděpodobnosti, že vytáhneme 10, 9, 8, 7, 6, 5, 4, 3, 2, 1 až žádnou kouli červenou při deseti tazích, čili pravděpodobnosti pro všechny hodnoty  $m$  od  $n$  až do nuly, dostaneme theoretické rozložení četností pravděpodobností čili t. zv. rozložení binomické.

**Poznámka.** Toto pravidlo se nazývá též zákon binomický, poněvadž hodnoty pravděpodobností po sobě následující představují po sobě jdoucí členy rozvoje binomu  $(p + q)^n$ .

Tabulka pravděpodobností vypočtených pro příklad č. 1.

Výskyt jevu $A$	Odchylka od normálního případu, t. j. 5 koulí červených a 5 koulí černých v serii deseti tahů		Pravděpodobnost výskytu $A$
10 červených koulí		+ 5	0,0009765625
9 „ „		+ 4	0,0097656250
8 „ „		+ 3	0,0439453125
7 „ „		+ 2	0,1171875000
6 „ „		+ 1	0,2050781250
5 „ „		0	0,2460937500
4 „ „		— 1	0,2050781250
3 „ „		— 2	0,1171875000
2 „ „		— 3	0,0439453125
1 „ „		— 4	0,0097656250
0 „ „		— 5	0,0009765625
			<hr/> 1,0000000000

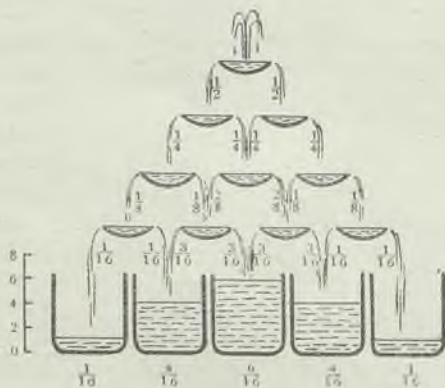
Binomické rozložení pravděpodobností ukazuje v tomto případě obraz zcela symetrický. V případech však, že  $p$  není rovno  $q$ , je rozložení a symetrické, při čemž se však stává čím souměrnější, čím více serie vzrůstají.

Vezmeme-li pravděpodobnost, že vytáhneme pět koulí červených a pět koulí černých, za případ normální, takový, jaký očekáváme podle obecné pravděpodobnosti  $\left(p = q = \frac{1}{2}\right)$ , potom značí nám ostatní případy odchylky od tohoto případu normálního, jak nahoře v tabulce je vyznačeno. Pravděpodobnost jednotlivých odchylek od případu normálního je tím menší, čím je odchylka větší.

Kdybychom si znázornili toto rozložení odchylek graficky, na př. pro 100.000 případů, násobíce 100.000 vypočtené pravděpodobnosti, obdrželi bychom zvláštní křivku jednovrcholovou a souměrnou, zvanou křivkou Gaussovou (pro uvedený případ, že  $p = q = \frac{1}{2}$ ).

Křivka Gaussova je ideálním obrazem rozdělení četností statistického souboru. Křivka tato je zcela souměrná, takže hodnoty po obou stranách osy  $y$  ležící se liší jen znaménkem, dále má tu vlastnost, že klesá rychle k ose úsečkové  $x$ , takže pravděpodobnost odchylky se zmenšuje v poměru k velikosti odchylky, a konečně se blíží stále k ose  $x$  (je asymptotická k ose  $x$ ).

**Poznámka.** Rozdělení odchylek podle zákona Gaussova lze znázorniti též mechanicky přístroji Galtonovým a Pearsonovým, vyobraz. u Yule-a, cit. český překlad, str. 309, nebo římskou studní, kde voda se záměrně rozděluje vždy na dva a další dva praménky podle schématu Schwarzova (A. Schwarz: Über den Umgang mit Zahlen, Mnichov 1943, str. 35).





**Římská studna.** Schema k demonstraci binomického rozložení. Voda teče z misky nejvýše položené dvěma stejnými pramenky do misek spodních, z těch opět stejnými pramenky do níže položených atd. Ve třetí řadě misek vytékající voda se spojuje,  $\frac{1}{8}$  a  $\frac{2}{8}$  na  $\frac{3}{8}$  množství vody, a ty se opět dělí na  $\frac{3}{16}$  a  $\frac{3}{16}$ . Množství vody v nádobách ve spodní řadě je rozděleno podle binomického rozložení.

## II. Pravděpodobnost theoretická a empirická.

V příkladech výše uvedených šlo vždy o pravděpodobnost předem uměle sestrojenou, matematickou, t. j. takovou, která se dá vypočítati předem na základě znalosti případu a všech jeho možností. Toto theoreticky vypočtené rozložení pravděpodobností lze pak srovnávati se skutečnými výsledky výše uvedených her náhodných, sestavených ve statistický soubor.

Avšak pravidelně neznáme u souborů statistických předem oněch možností, které mohou nastati, nemůžeme tudíž ani vypočítati pravděpodobnost onoho jevu, který nás zajímá. Postup, který sledujeme při zkoumání těchto jevů, je tedy jaksi opačný: z velkého pozorovaného souboru těchto jevů, u kterých se hledaná událost (znak) vyskytla, usuzujeme na pravděpodobnost jejich podle získané zkušenosti, empiricky. Schematicky lze to vyjádřiti vzorcem: učinili-li jsme počet  $n$  pozorování nějakého souboru a při nich v  $m$  případech vyskytl se onen jev, který nás zajímá, a jehož pravděpodobnost bychom chtěli vyšetřiti, vyjadřuje nám poměr  $\frac{m}{n}$  četnost (frekvenci) tohoto jevu uvnitř onoho souboru. Při velkém počtu pozorování, rozsáhlém souboru, předpokládáme, že blíží se tato četnost pravděpodobnosti theoretické, t. j. takové, jakou bychom předem theoreticky očekávali, kdyby nám byly veškeré podmínky výskytu takového jevu známy. Toto pravidlo o vztahu velkého počtu pozorování nějakého jevu k theoretické pravděpodobnosti se nazývá zákonem velkých čísel.

**Poznámka.** Název „zákon velkých čísel“ pochází od *Poissona* (*Recherches sur la Probabilité des Jugements en matière criminelle et en matière civile*, 1837).

Sama zkušenost vedla k poznání, že při velkém počtu pozorování se objevují pravidelnosti ve složení souborů statistických, které nelze zjistiti, jestliže vybereme jen jednotlivé jevy samy o sobě nebo pozorujeme pouze malý počet jich. Lidé dosahují věku velmi různého; někdo zemře v dětském věku, jiný v květu mládí, opět jiný dožije se vysokého stáří. Konkretní životní osudy samy o sobě pozorovány zdají se býti zcela nepravidelné, jak říkáme náhodné, a nemůžeme z nich vyvozovati žádné úsudky. A přece životní pojišťovny, jejichž obchody zakládají se na pojišťování

pro případ úmrtí, na základě výpočtů určují pojistné prémie a ze svých obchodů vykazují každoročně pravidelné zisky.

Tuto pravidelnost, známou též pod klasickým označením zákon velkých čísel, přirovnal *Quételet* k pozorování kružnice křídou na tabuli narýsované. Pozorujeme-li ji zblízka, rozpadá se kružnice na veliké množství samostatných, zcela nepravidelně rozptýlených bodů, křivka sama se však zcela ztrácí. Teprve odstoupíme-li poněkud, dostáváme obraz souvislé křivky, body splývají v jedno a ztrácejí se, a objevuje se jejich symetrické, pravidelné uspořádání, vidíme kružnici.

Právě tato poměrně velká pravidelnost a stálost, která se jeví v soubozech jevů společenských, má základní význam pro různá společenská a hospodářská zařízení. Na základě dlouhé zkušenosti jsou založena všechna společenská zařízení na předpokladu jisté pravidelnosti a stálosti všech jevů. Ústavy veřejné, školy, nemocnice, trestnice, ústavy zaopatřovací vesměs jsou zařízení jen pro jistý normální, předpokládaný počet osob, který se rok od roku pouze málo mění, podobně i život hospodářský je vypočten na jistou pravidelnost: obchodník počítá s určitým stálým počtem zákazníků a podle toho nakupuje a udržuje své zásoby, likvidita peněžních ústavů je založena na podobné pravidelnosti, státní rozpočet vychází rovněž z předpokladů stálosti příjmů a vydání v budoucím roce.

Tyto jevy jsou namnoze povahy biologické, na př. četné kulturní formy a právní instituce založeny jsou na trvalém předpokladu, že poměr pohlaví u novorozeneckých dětí zůstává stálý a drží se přibližně v rovnováze s malým přebytkem chlapců.

**Příklad č. 2.** *Poměr počtu narozených chlapců k počtu narozených děvčat v Československu.*

Na 1000 narozených děvčat připadalo chlapců (zaokrouhleno na celá čísla)

r. 1920	1079	r. 1929	1074
1921	1077	1930	1062
1922	1078	1931	1068
1923	1071	1932	1065
1924	1066	1933	1064
1925	1071	1934	1068
1926	1066	1935	1062
1927	1061	1936	1072
1928	1064		

**Poznámka.** Veškerá data v tomto příkladě i v ostatních příkladech uvedená vztahují se na území Československé republiky z počátku r. 1938.

Zákon velkých čísel je potvrzován četnými pokusy konanými u t. zv. náhodných her, na př. při opakovaných tazích koulí z urny se stejným počtem černých a bílých koulí, kam se vytažené koule opět vracejí, kde tedy očekávaná pravděpodobnost pro každý tah pro každou barvu je  $\frac{1}{2}$ .

Při 10.000 opakovaných tahů došel *Westergaard* k poměru tažených koulí podle barvy 50,1% a 49,9%. Tyto tahy analysoval pak blíže tím způsobem, že konal je v seriích po stu tazích, které pak vždy k předcházejícím přičítal, a zkoumal, jak se blíží očekávané pravděpodobnosti. Výsledky potvrdily, že poměr obou barev tím více se blížil očekávanému poměru  $\frac{1}{2}$  pro každou barvu, čím bylo provedeno více tahů, čili čím více serií k sobě přičítal, tím byly poměrné odchylky od očekávaného výsledku menší.

Sestavíme-li výsledky takových pozorování, jako byla *Westergaardova*, zaznamenaných podle serií, v tabulku seřazenou podle velikosti odchylek od očekávaného (normálního) výsledku (čili sestavíme-li z výsledků pozorování statistický soubor a tabulku rozložení četností), můžeme zjistiti, kolikrát se která odchylka vyskytuje, čili zjistiti četnosti těchto odchylek. Aritmetický průměr tohoto souboru rovná se  $A = sp$  a směrodatná odchylka souboru rovná se  $\sigma_0 = \sqrt{spq}$  (jestliže opět označíme písmenem  $p$  očekávanou pravděpodobnost jevu  $A$ , písmenem  $q$  očekávanou pravděpodobnost jevu protilehlého a vzájemně se vylučujícího  $B$  a písmenem  $s$  celkový počet pozorování v serii: srv. dále kap. o statistických charakteristikách).

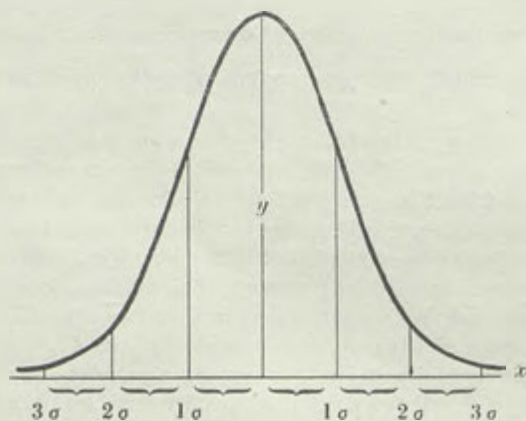
Poměrná směrodatná odchylka, t. j. směrodatná odchylka četností uvedená v poměr k počtu pozorování  $s$  rovná se pak

$$\frac{\sigma_0}{s} = \sqrt{\frac{spq}{s}} = \sqrt{\frac{pq}{s}}.$$

Z tohoto vzorce plyne, že poměrná odchylka (očekávaná) je tím menší, čím větší jest  $s$  počet pozorování. Výsledek pozorování bude se tedy tím více blížíti očekávané pravděpodobnosti, čím je větší počet pozorovaných případů, avšak nikoli úměrně s počtem pozorování, nýbrž toliko s jeho druhou odmocninou. Zvýšíme-li tedy počet pozorování čtyřikrát, přiblíží se skutečné rozložení odchylek očekávané pravděpodobnosti pouze dvakrát, stále za předpokladu, že se neporušuje stejnorodost souboru. V seskupení odchylek kolem aritmetického průměru se objevuje velká pravidelnost, odchylek ubývá tím více, čím jsou větší a přibývají



souměrně v obou směrech, čím jsou menší. V mezích jedné směrodatné odchylky  $1\sigma$  leží 68,27% všech případů, v mezích dvojnásobné směrodatné odchylky  $2\sigma$  95,45% případů a v mezích trojnásobné odchylky  $3\sigma$  99,73% všech případů, čili zhruba skoro všechny případy. Odtud je odvozeno pro statistickou praxi důležité pravidlo, že statistické hodnoty, poměry a koeficienty považujeme jen tehdy za ukazatele spolehlivé, jestliže převyšují aspoň trojnásobně svou theoretickou směrodatnou odchylku, jinak že se pohybují v rámci náhodných odchylek (v. obrazec *Gaussovy* křivky a rozdělení její plochy).



Všeobecně lze nazvat zákonem velkých čísel úkaz, že statistické soubory vykazují teprve od určité velikosti svou charakteristickou strukturu, která není zřejmá u jednotek nebo u malých skupin, z nichž se soubor skládá.

Matematicky lze vyjádřit zákon velkých čísel teorémem *Bernoulliho* pro případy, kde pravděpodobnost události zůstává nezměněna při přechodu od jednoho jevu ke druhému (na př. vytahují-li se koule z urny a zase se do ní vracejí, takže poměr barev zůstává při každém tahu týž), pro ostatní případy byl pak doplněn *Poissonem*.

Tento teorém dá se v podstatě objasniti slovy:

Maximum theoretické pravděpodobnosti odpovídá maximu empirické, skutečné pravděpodobnosti, takže při dostatečně velkém počtu pozorování lze očekávati, že jev o největší theoretické pravděpodobnosti se také ve skutečnosti vyskytne nejčastěji. Při počtu pozorování vzrůsta-



jícím (theoreticky) do nekonečna, blíží se rozdíl mezi theoretickou pravděpodobností a skutečným výskytem jevu nule.

Vzorec *Laplaceův* vyjadřuje pravděpodobnost, že poměrná četnost (skutečné vyskytování) události se neodchýlí od matematické pravděpodobnosti této události více než o určitou veličinu, čili že se bude odchylka od matematické pravděpodobnosti pohybovat v mezích daných touto veličinou.

Označíme-li tuto veličinu písmenem  $\Delta$ , je dána tato pravděpodobnost funkcí výrazu

$$u = \frac{\Delta}{\sqrt{\frac{2pq}{s}}}$$

kde ostatní písmena mají stejný význam jako v předcházejících vzorcích. Funkci tohoto výrazu lze přibližně vyjádřit integrálem

$$\Phi(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2} dt$$

( $\pi$  značí Ludolfovo číslo 3,14159,  $e$  základ přirozených logaritmů 2,71828,

$t = \frac{x}{\sqrt{\frac{2pq}{s}}}$ , při čemž  $x$  je hodnota úsečky každé odchylky, měřená od středu

symetrického rozložení).

**Poznámka.** O funkčním poměru dvou veličin mluvíme tehdy, jestliže každé hodnotě veličiny jedné odpovídá určitá hodnota veličiny druhé. Počet infinitesimální jedná o veličinách proměnných: v diferenciálním počtu určujeme ze známé závislosti dvou veličin proměnných nekonečně malou změnu jedné, odpovídající nekonečně malé změně druhé veličiny. V počtu integrálním obráceně hledáme zákon, spojující obě proměnné veličiny, ze známých sobě odpovídajících nekonečně malých změn jejich (srv. *Vojtěch: Základy matematiky*, Praha 1922, I, str. 5).

Pro jednotlivé hodnoty veličiny  $u$  je možno použití tabulek, aniž by bylo třeba vypočítávat integrál pro každou hodnotu. Obsahuje je na př. cit. kniha *Kohnova*, str. 175. Pro přehled postačí, uvědomíme-li si, že hodnota  $\Phi(u)$  pro případ, že  $u = 1$ ,  $\Phi(u) = 0,843$ , když  $u = 2$ ,  $\Phi(u) = 0,995$ , když  $u = 3$ ,  $\Phi(u) = 0,99998$ , čili blíží se prakticky jedné celé.

Rozložení *Laplaceova* integrálu tvoří v grafickém znázornění křivku *Gaussovu*.

**Poznámka.** Z výrazu  $u = \frac{\Delta}{\sqrt{\frac{2pq}{s}}}$  lze odvodit jiné měřítko pravděpodobnosti,

t. zv. modul. Z uvedeného vzorce vyplývá, že odchylka  $\Delta = u \sqrt{\frac{2pq}{s}}$ ; položíme-li  $u = 1$ , dostaneme výraz  $\sqrt{\frac{2pq}{s}}$ , který se označuje jako modul a je mírou mezních odchylek poměrných četností očekávaných událostí od jejich theoretické pravděpodobnosti.

Je to vlastně poměrná odchylka  $\sigma_0 = \sqrt{\frac{2pq}{s}}$ , násobená  $\sqrt{2}$  (srov. str. 27 a 58 sl.).

**Příklad č. 3.** Máme li v urně čtyři koule černé a šest bílých, pravděpodobnost (theoretická), že vytáhneme kouli černou  $p = \frac{4}{10} = 0,4$ , pravděpodobnost, že vytáhneme bílou  $q = \frac{6}{10} = 0,6$ .

Modul umožňuje nám zjistiti pravděpodobnost, že při stu opakovaných tazích poměrné vyskytování černé koule  $\left(\frac{m}{n}\right)$  bude se pohybovati v mezích daných modulem.

Modul v tomto případě se rovná  $= \sqrt{\frac{2 \times 0,4 \times 0,6}{100}} = 0,069$ ,

tudíž meze dány jsou hodnotami  $0,4 + 0,069$  a  $0,4 - 0,069$ , čili tážeme se na pravděpodobnost, že při stu tahů bude se tah černé koule pohybovati v mezích  $0,469$  a  $0,331$ .

Tato pravděpodobnost je dána v tabulce hodnot  $\Phi(u)$  pro případ zvolený, že  $u = 1$ , číslem 0,843. Ze sta tahů bude se tedy přibližně 84 pohybovati v uvedených mezích, ostatní tahy (16) budou tyto meze přestupovati.

Spolehlivost předpovědi pravděpodobného výsledku můžeme zvýšiti, vezmeme-li modul dvojnásobný nebo trojnásobný (čímž ovšem se modul dvojnásobně nebo trojnásobně zvětšuje a tím i meze, v nichž se budou tahy pohybovati), jakož i zvýšením počtu pozorování v serii (číslo s).

Vykonáme-li v uvedeném příkladě deset tisíc tahů, bude se modul rovnati 0,0069, meze budou pak dány hodnotami  $0,4 \pm 0,0069$ , čili 0,4069 a 0,3931, čili jsou daleko užší než v případě prvním; přesnost výsledku vzrůstá úměrně s druhou odmocninou počtu pozorování.

**Literatura ke kapitole 3—4.** *St. Kohn*, Základy theorie statistické metody, *W. Wundt* Logik I—III, 4. vyd. Stuttgart 1919—1921, *H. Poincaré*, La science et l'hypothèse, Paříž 1920, *U. Yule*, Úvod do theorie statistiky, *A. A. Čuprov*, Očerki po teorii statistiki, Moskva 1910, *týž* Grundbegriffe und Grundprobleme der Korrelationstheorie, Lipsko 1925.

## Kapitola 5.

# TECHNIKA STATISTICKÉHO ŠETŘENÍ.

Číselné údaje statistické zvané statistická data nebo statistický materiál získáváme hromadným pozorováním, které stručně označujeme jako statistické šetření. Technikou statistického šetření rozumíme pak způsob a postup, jakým soubor zachycujeme. Statistická šetření souborů skládajících se z jevů trvalých bývají označována jako sčítání nebo soupis (census, na př. sčítání lidu, sčítání živnostenských závodů, sčítání ovocných stromů). Jimi se zjišťuje stav souboru v určitém momentu časovém (na př. stav obyvatelstva 1. prosince určitého roku). Soubory, skládající se z jevů okamžitých (událostí), zachycujeme statisticky trvalým (neustálým) pozorováním (registrováním) vyskytnuvších se případů (událostí).

Techniku statistického šetření můžeme si rozdělit za účelem lepšího přehledu na tři hlavní stadia:

1. přípravu šetření,
2. vlastní šetření,
3. zpracování, úpravu a publikaci získaného materiálu.

### I. Příprava statistického šetření.

Přípravné práce ke statistickému šetření vyžadují nejprve, aby byl stanoven plán, podle kterého se má šetření konati. Je třeba určit, který soubor chceme znáti a co na něm chceme znáti. Soubor, který chceme podrobiti statistickému šetření, musíme nejprve logicky pevně vymeziti,



definovati. Při tom je statistik obyčejně vázán pojmy, vytvořenými jinými vědami, na př. národním hospodářstvím, jde-li o soubor jevů hospodářských, nebo právními normami. Jestliže však pojmy, které takto jsou poskytovány, nejsou zcela jednoznačné, je třeba, aby si je statistik vymezil samostatně tak, aby bylo možno zjistiti bezpečně, zda v konkrétním případě je jev dán čili nic. Na př. pojem kupní síla peněz, národní důchod a pod.

Soubor musí býti vymezen jednak časově, jednak místně. Pokud jde o čas, musí býti stanovena kritická doba, kdy se má konati šetření souborů jevů trvalých, nebo musí býti stanoveno časové období, interval, průběhem kterého se má soubor pozorovati, při souborech jevů okamžitých. Pokud jde o místo, musí býti stanoveno území, na které se má šetření vztahovati.

Věcné vymezení souboru děje se prakticky vymezením jednotky statistického šetření, ze kterých se pozorovaný soubor skládá a na které jest jej třeba rozložit.

Jednotkou šetření, ze kterých se skládá soubor, hromadný jev, je na př. při sčítání lidu individuum (jednotlivý obyvatel), při sčítání živnostenských závodů technické a výrobní zařízení, které jest definováno jako živnostenský závod. Někdy může býti při jednom šetření zjišťována jednotka dvojí, na př. při trestní statistice je jednotkou trestný čin a zároveň pachatel trestného činu, v tomto případě ovšem jediným šetřením zjišťujeme současně dva soubory, soubor delinkventů a soubor trestných činů, oba sice úzce spolu souvisící, ale nikoli totožné, neboť jeden delikvent se může dopustiti více trestných činů, a naopak na jednom trestném činu může míti podíl více delinkventů.

Z povahy statistiky a statistického šetření vyplývá, že jednotka šetření má vyhovovati těmto požadavkům: z jednotek musí býti složen soubor pozorovaný, jednotka musí se dáti vyjádřiti číselně a musí býti tak definována, aby ji bylo možno bezpečně zjistiti.

Definice jednotky šetření vychází ovšem zase z věcné povahy předmětu šetření, musí přihlížeti k účelu šetření a pojmům, vytvořeným příslušnými vědami, ale při tom se zároveň říditi požadavky statistické metody. Na př. při sčítání lidu je jednotkou šetření ovšem individuum, ale třeba blíže vymeziti, které, zda to, jež jest v kritické době na státním území přítomno, nebo to, které má tam stálé bydliště (podle toho zjistíme pak soubor obyvatelstva přítomného nebo obyvatelstva trvale bydlícího na státním území); při sčítání domů je třeba vymeziti, je-li jednotkou šetření dům ve smyslu jednotky administrativní nebo stavebně-technické,

rozřešiti otázku, kam zařaditi obývané baráky a nouzové kolonie bez popisných čísel, weekendové chaty a pod., avšak koná-li se šetření za účelem fiskálním, je třeba položit za základ definice jednotky šetření definici zákona o dani domovní.

Podobně jako jednotku šetření je třeba definovati též znaky, t. j. vlastnosti jednotek, které chceme zjistiti. Podle povahy znaků, jsou-li kvalitativní nebo kvantitativní, přetržité nebo nepřetržité (spojité), jest třeba zaříditi též otázky po nich. Definice znaků se řídí opět pojmy vytvořenými příslušnou disciplinou, ale se zřetelem na zásady statistické metody a účel šetření. Na př. znak národnosti při sčítání lidu definuje se jako příslušnost kmenová, jejímž hlavním znakem je zpravidla mateřská řeč, ale vzhledem k účelu šetření připouští se výjimky, na př. pro židy, náboženské vyznání se definuje jako příslušnost k některé konfesi státem uznané, ale je třeba zařaditi i znak bez vyznání nebo příslušnosti k některé církvi sice právním řádem neuznané, ale skutečně existující, jako církvi metodistické a pod.

Pro každé statistické šetření souborů jevů trvalých, tedy t. zv. sčítání, je třeba zvoliti vhodnou dobu, ve které se šetření má konati, t. zv. dobu kritickou. Kritická doba řídí se povahou souboru a opět účelem šetření; přihlíží se obvykle k tomu, aby v ní právě byl celý soubor v klidu a ve stavu „normálním“; na př. sčítání lidu koná se v měsících zimních (u nás 15. února, poslední 1. prosince), kdy obyvatelstvo podle zkušenosti zdržuje se nejvíce ve svém obvyklém bydlišti; sčítání zemědělských závodů na jaře po skončení jarních prací zemědělských a pod. Aby bylo možno výsledky sčítání srovnávati s budoucími, opakují se sčítání pokud možno v pravidelných obdobích, na př. pětiletých, desítiletých; mluví se pak o statistických šetřeních periodických.

Při šetření souborů jevů okamžitých (událostí) koná se statistické šetření tím způsobem, že se stále nepřetržitě pozorují a zaznamenávají jednotky takového souboru; na př. sňatky, úmrtí, porody neustále se zaznamenávají v matrikách, přechod zboží přes celní hranice se neustále zaznamenává jako běžné, nepřetržité statistické šetření zahraničního obchodu.

Podle uvedených směrnic se pak sestavuje dotazníkový formulář, obsahující otázky, potřebné ke statistickému šetření. Při sestavování dotazníku je třeba uvážiti nejen rozsah šetření, který je k danému účelu šetření potřebný, nýbrž i posouditi podrobně, jaké otázky mohou býti kladeny, aby bylo možno očekávati na ně správné odpovědi, které otázky jsou nejúčelnější, jak mají býti otázky formulovány, aby byly srozumí-

telné, kdo má být dotazován; dále předem rozhodnouti, co má a může být zpracováno z výsledků šetření. Základní směrnice pro každé šetření a sestavování dotazníku jsou: dotazovati se pouze na věci pro vyšetření souboru nutné, netázati se však na věci zbytečné, na takové, u nichž nelze odpovědi kontrolovati (na př. věci čistě subjektivní povahy), na věci takové, které pravděpodobně nebude lze zpracovávat, na věci soukromého nebo rodinného života, na takové, při nichž odpověď se zdá býti choulostivá nebo směšná.

Další podrobnosti plánu se týkají organizace šetření, které statistické orgány mají šetření konati, doby, kdy je mají konati a způsobu, jakým je mají konati.

Všechny tyto směrnice a požadavky lze však uplatniti toliko při *samostatných* šetřeních statistických, která poskytují t. zv. primární statistický materiál, to jest tam, kde pozorujeme jednotky šetření, ať jde o jevy trvalé či okamžité, jediné za účelem statistickým, naproti tomu při t. zv. *sekundárním* statistickém materiálu je statistik vázán dodaným materiálem a zpracování jest omezeno na údaje získané k jiným účelům než statistickým, obyčejně správním. U statistického materiálu sekundárního je tedy třeba přizpůsobiti i jednotky šetření pojmům, které vznikají z norem správních. Tak na př. statistika porodů, úmrtí, sňatků zakládá se na materiálu sekundárním, poněvadž se tyto jevy populační znamenávají v matrikách za účely správními; sbírání materiálu řídí se tedy předpisy o vedení matrik.

Sekundární statistický materiál může býti doplněn primárním, tím, že se kladou *doplňující otázky statistické*, na př. právě při statistice pohybu obyvatelstva správcové matrik doplňují záznamy zřízené k účelům správním některými otázkami povahy primárního statistického šetření, na př. otázkou po národnosti snoubenců, jejich budoucím společném bydlišti atd.

Dotazníky jsou buď individuální, pro každou sčítanou jednotku zvláštní, na př. při sčítání zemědělských a živnostenských závodů jest pro každý závod určen zvláštní dotazník, nebo *hromadné, kolektivní*, společné pro více jednotek šetření, nějakým způsobem organicky seskupených, na př. sčítací arch pro domácnost při sčítání lidí.

Vedle vlastních otázek obsahují dotazníky pravidelně ještě potřebné sankce trestní, t. j. hrozbu trestem na nezodpovědění otázek nebo nesprávné odpovědi, ujištění o ochraně tajemství poskytnutých údajů, identifikaci dotazníku běžným (pořadovým) číslem nebo jiným poznávacím znamením, dále příklady, jak odpovídati na otázky, vysvětlivky k jednotlivým otázkám, poučení o účelu šetření a j.



## II. Statistické šetření.

Na tomto místě máme na zřeteli pouze úplné šetření souboru místně a časově vymezeného. Šetření samo děje se buď přímým pozorováním jednotek souboru, ať jsou to osoby nebo věci, nebo dotazováním osob, které potřebné údaje mohou poskytnouti a byly v přípravném plánu k tomu úkolu vybrány a určeny.

První případ, *objektivní pozorování* jednotek šetření, je ve statistice sociální řídký, na př. statistika poštovních zásilek, odesílaných v určitý den, nebo pouliční frekvence, zjišťovaná v některé čtvrti města, za to často se vyskytuje ve statistice anthropometrické, biometrické, při technických šetřeních a j.

Pravidelný případ šetření ve statistice sociální jest způsob druhý, *vyšetřování souboru dotazováním osob*, které k tomu byly zvoleny. Jsou to podle povahy souboru a účelu šetření osoby, které mají největší znalosti o něm a od kterých možno očekávat i též správnou odpověď na kladené otázky. Na př. při sčítání lidu přednostové domácnosti, při statistice zahraničního obchodu příjemce zboží, při dovozu odesílatel, nebo vývoze při vývozu, při sčítání zemědělských podniků vlastníků nebo nájemce podniku a pod.

Dotazy se dějí nejčastěji písemně, ve formě dotazníku, který se těmto osobám předkládá k vyplnění, někdy ústně, kde odpovědi zapisuje orgán statistické služby. Oba způsoby mohou být též kombinovány, což se děje hlavně za účelem kontroly spolehlivosti a úplnosti údajů. Obvykle se postupuje tak, že dotazníky, písemně vyplněné osobami poskytujícími údaje, jsou pak u nich orgány statistické služby sbírány a současně kontrolovány.

Podle subjektu, který statistické šetření podniká, lze rozeznávat soukromá a veřejná (úřední) statistická šetření. Sama povaha statistiky přináší s sebou, že úzce souvisí s veřejnou správou. Byla již učiněna zmínka o sekundárním statistickém materiálu, který je vlastně produktem státní a veřejné správy. Státní a veřejná správa, která má stálou tendenci rozšiřovati svoji působnost, přináší a zaznamenává množství údajů tvořících statistický materiál (na př. statistika demografická, trestní, veřejného pojištění, monopolů, veřejných financí atd.). Tento materiál poskytovaný státní správou, je též nejstarším historickým statistickým materiálem. Naopak opět veřejná správa potřebuje ke své činnosti znalosti poměrů, které spravuje, a tak sama vyvolává nová statistická šetření. Vedle této souvislosti mezi veřejnou správou a stati-

stickými šetřeními jsou i jiné okolnosti příznivé veřejným (úředním) statistickým šetřením: donucovací moc státu, často opírající se o generelní klausuli zákonnou, může přinutiti občany, aby poskytovali požadované údaje pod sankcí trestů, což je u statistiky soukromé nemožné. Rozsáhlý správní aparát státu a jiných korporací veřejnoprávních může býti postaven do služeb statistického šetření, což při velkém rozsahu četných statistických šetření je podmínkou pro jejich zdar. Stát může si opatřiti lépe též potřebné odborné síly i finanční prostředky, neboť statistická šetření velkého rozsahu jsou samozřejmě i velmi nákladná. Všechny tyto okolnosti, nehledíme-li ani ke skutečnosti, že stát zasahuje v poslední době stále pronikavěji do sociálních poměrů, vedou k tomu, že státní příp. veřejná statistika zatlačuje čím dále tím více statistiku soukromou, která se nyní omezuje na šetření k účelům vědeckým, na př. lékařským, anthropometrickým, biometrickým, dále technickým, pojišťovacím a j. Konání statistických šetření je nyní vázáno na schválení Státním úřadem statistickým podle vládní nařízení č. 186/1940 Sb.

Ovšem též veřejná statistická šetření mají některé nevýhody. Je to častá nedůvěra ke státním orgánům, obava před zdaněním nebo jinými nepříznivými důsledky poskytnutých údajů (statistických údajů získaných protektorátní statistikou bylo přímo užíváno k předpisu povinných dodávek, na př. ovoce!). Též administrativní rozdělení státního území neodpovídá vždy skutečným poměrům zeměpisným, demografickým, hospodářským a j., které jsou vyšetřovány, a finanční situace státu omezuje často rozsah šetření.

V Československu byla přijata jako ve většině států moderní zásada soustředění statistických šetření, ať svou povahou spadá materiál statistický do kteréhokoli oboru státní správy. Veškerá statistická šetření dějí se až na malé výjimky zásadně u Státního úřadu statistického, zvláště k tomu účelu zřízeného.

Organisace statistické služby státní se zakládá na zvláštním zákoně ze dne 28. ledna 1919 č. 49 Sb. Na Slovensku byl zřízen Štátny plánovací a štatistický úrad nařízením Slovenské národní rady ze dne 26. května 1945 č. 48 sb. S. N. R. Jednota statistické služby u obou těchto úřadů byla upravena dohodou vlády a Slovenské národní rady z dubna 1946.

Vedle státního úřadu statistického provádějí statistická šetření povahy veřejné ještě městské statistické kanceláře hlav. města Prahy a některých větších měst. Jejich pravomoc není až dosud upravena zvláštním zákonem. Úřady ty konají některá šetření místní povahy, zpracují a uveřejňují samostatně jejich výsledky, stejně jako výsledky státních

šetření pro příslušnou obec, jinak spolupracují se Státním úřadem statistickým.

Jako podřízené výkonné orgány fungují při některých statistických šetřeních různé orgány správy vnitřní, finanční i soudní, podle povahy šetření, na př. politické úřady a obecní úřady při sčítání lidu, celní úřady při statistice zahraničního obchodu, soudní kanceláře při statistice soudní atd. Pro velká sčítání se zjednávají vedle toho i statistické orgány mimořádné, jako sčítací komisaři při sčítání lidu.

### III. Zpracování a uveřejnění statistických dat.

Statistický materiál, který vznikl jako výsledek šetření, musí být před konečným zpracováním kontrolován, je-li úplný a správný. To se děje buď u podřízených orgánů, které šetření konají, nebo u úřadu, který má provést konečné zpracování, jestli sám šetření koná. Hlavní účel této kontroly jest zabránit tomu, aby nebyly části pozorovaného souboru vynechány, na př. domy při sčítání lidu.

Statistický materiál v původním stavu, t. j. individuální údaje, zpracují se zásadně nejprve sčítáním, a to podle skupin, t. j. jednotky se současně třídí podle různých znaků, místních, časových, věcných.

Při šetřeních většího rozsahu, t. j. kde se zjišťuje větší počet znaků, používá se k dalšímu zpracování t. zv. *sčítacích lístků neboli štítků*, t. j. přesně stejných kartonů, na které se přenášejí zjištěné znaky. Každý šetřený případ obdrží jeden štítek (na př. při sčítání lidu každý obyvatel = jednotka šetření = štítek), na který se zaznamenává, který znak, v které obměně nebo velikosti, se u něho vyskytl. Toto zaznamenávání se děje obvykle postupem zvaným *dírkování*, t. j. na pořízené sestavě číslic se vyznačí podle umluveného klíče proražením na určitém místě znak a jeho stupně (nebo obměny); obvykle děje se tak mechanicky t. zv. *dírkovacím strojem*. Kde je to možno, lze použít též již dotazníkového formuláře, vyhotoveného v podobě štítku, což znamená, že se uspoří přenášení znaků ze sčítacího archu na zvláštní štítek.

Sčítání samo se může dít různým způsobem:

1. *čárkováním*, opakujícím se tolikrát, kolikrát se který znak vyskytuje (na př. pohlaví mužské), každý ze sčítaných případů, kde se vyskytuje, vyznačí se čárkou, které pak za účelem lehčího sčítání a kontroly se seskupují v různé obrazce, obsahující určitý počet jednotek, na př. po pěti:



a pod.;



## Příklad č. 4.

Vzor štítku perforovaného, používaného při statistice zahraničního obchodu  
(perforovaná místa jsou tmavě vyznačena).

Rok	12 tř. 11 má	12 tř. 11 dub.	Řadové číslo						Komorní obvod			Statist. číslo						Váha kg — g						Hodnota v Kčs						Země zaslky, převodu	Země vý- rob, měřeni	Kusy																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27			28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																									
30	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

STATIST. ZAHR. OBCHODU-DVOVOZ-VÝVOZ.

2. lepením známek z útržkového bloku. Pro každý znak jest určena zvláštní známka, pro každý případ se nalepi známka s běžným číslem. Přitom jest možno i sčítati zároveň různé kombinace znaků, na př. kombinaci pohlaví mužského, národnosti německé, náboženství římsko-katolického, pro každou kombinaci musí býti určena známka určité odlišné barvy;

3. skládáním sčítacích lístků (štítků). V tomto případě musí již znaky býti vyznačeny nebo převedeny na štítky. Štítky se pokládají podle stejného znaku na sebe a složené hromádky se pak sčítají.

Daleko častěji, a to zvláště při všech větších statistických šetřeních, používá se však mechanického zpracování štítků zvláštními, k tomu účelu konstruovanými stroji statistickými. Jak již bylo uvedeno, znaky se převádějí podle zvláštního klíče na systém číselný, kde se pak vyznačují pomocí dírkovacích strojů na štítcích dírkami (perforacemi). Štítky tímto způsobem dírkované se pak automaticky třídí a skupiny sčítají třídícími stroji, zakládajícími se v podstatě na tom, že elektrický proud prochází volně dírkami, jest však přerušován kartony štítků na těch místech, kde nejsou dírkovány. Skupiny perforovaných štítků, vzniklé tříděním, se pak sčítají stroji tabulačními. Stroje tyto (vynalezené Holerithem, nyní nejvíce používány jsou velmi zdokonalené stroje Powersovy) jsou poháněny elektricky a umožňují velmi rychlé zpracování velikého množství statistických údajů, jsou tedy velmi vhodné zejména pro velká statistická šetření a vůbec pro systém centrálního zpracování statistických dat. Zpracování takové při rozsáhlé dělbě práce má ráz tovární velkovýroby.

Zpracování statistického materiálu nazývá se též francouzským názvem dépouillement.

Výsledky sčítání se vyznačují do pomocných tabulek pracovních, ze kterých se pak sestavují tabulky konečné.

Po stránce organizační děje se zpracování samo buď podle zásady centralisační, takže veškerý statistický materiál se dodává jedinému centrálnímu úřadu, kde se pak jednotně zpracuje (jako je tomu u nás), nebo decentralisační, kde se materiál zpracovává hned u orgánů statistických na místě šetření a pouze konečné zpracování se děje v úřadě centrálním. Výhody centralisační zásady, jednotný způsob zpracování, dělba práce, možnost opatření strojní nákladná zařízení a odborné síly, způsobují, že tato zásada v moderní statistické službě převládá; naproti tomu podrobná kontrola údajů působí při systému centralisačním někdy obtíže.

Zpracované výsledky statistických šetření sestavují se v tabulky, které pak slouží jako pramen dalšího hodnocení a vědeckého zpracování.

Tabulky jsou číselné přehledy výsledků statistických šetření, k nim se pak přidává text, především hlavička a legenda, které uvádějí pojmy, jež čísla dále vyjadřují. Dále se připojují poznámky o šetření, zejména vysvětlení, co je jednotkou souboru, jednotkou šetření, definice znaků atd.; k tomu přistupuje ještě obvykle popis šetření, výklad o použitých dotazníkových formulářích a pod. Konečně se obvykle připojuje výklad o výsledku šetření, ve formě všeobecné nebo podrobnější, případně mající již povahu vědeckého rozboru vyšetřeného materiálu.

Tabulky se rozdělují systémem kolmých a vodorovných přímek na sloupce a řádky, a umožňují výsledky uveřejňovati podle zjištěných znaků místních (na př. údaje pro jednotlivé obce, soudní okresy, politické okresy, země atd.), časových (na př. pro měsíce, roky), věcných (na př. druhy povolání obyvatelstva). Se zřetelem k nutné přehlednosti tabulek nedoporučuje se vyznačiti na jedné tabulce více než čtyřnásobnou kombinací znaků. Kombinací rozumíme rozdělení souboru současně podle dvou nebo více znaků, na př. obyvatelstva podle národnosti a povolání, nebo skládání jednotek podle různých znaků (věcných, místních, časových). Seskupováním jednotek podle kombinací znaků postupujeme pak při zkoumání závislosti mezi znaky.

Vedle čísel absolutních zařazují se v tabulky též čísla poměrná.

Výsledky statistických šetření uveřejňují statistické úřady v publikacích zvaných *pramenná díla*, která obsahují podrobné údaje šetřením získané, ať jde o sčítání nebo stálá statistická zaznamenávání za určitá období; vedle toho obsahují též tabulky s kombinacemi v různém rozsahu, grafická znázornění výsledků a příslušný text.

*Literatura ke kapitole 5. D. Krejčí Základy statistiky, 2. vyd.*



## STATISTICKÉ ZNAKY A ROZLOŽENÍ STATISTICKÝCH SOUBORŮ.

### I. Statistické znaky.

Účelem statistického pozorování je popisovati a analysoвати soubory hromadných jevů. K tomu je třeba znáti nejen rozsah souborů, nýbrž též znaky souborů, t. j. některé vlastnosti, které jevy blíže určují a rozlišují. Tyto znaky, které chceme vyšetřiti, dělí se podle své povahy na dvě hlavní skupiny:

1. znaky *kvalitativní*,
2. znaky *kvantitativní*.

Znaky kvalitativními nazýváme ony znaky, které vyjadřují nějakou vlastnost, jež se nedá měřiti, a při nichž při statistickém pozorování pouze zjišťujeme, zda se znak u pozorované jednotky vyskytuje čili nic. Na př. mužské pohlaví, náboženské vyznání římsko-katolické, povolání zemědělec při sčítání lidu; u každého vyšetřovaného obyvatele (jednotky) zjišťujeme, zda se tento znak u něho vyskytuje čili nic. Také znaky kvalitativní jsou měnlivé potud, že se vyskytují obměny znaku, buď dvě, na př. pohlaví mužské a ženské, t. zv. alternativní obměny znaku kvalitativního, nebo více obměn, t. zv. množné obměny nebo množné třídění znaku kvalitativního, na př. národnost česká, německá, polská, maďarská, ruská a jiná.

Znaky kvantitativní jsou ty znaky, které jsou obsaženy u každé jednotky souboru, ale které se od jednotky k jednotce mění a které se dají vyjádřiti různými hodnotami (kvantitativně) podle nějakého obvyklého

měřítka. Jsou to znaky, které se dají buď měřiti nebo sčítati. Na př. věk, který se vyskytuje u každého pozorovaného obyvatele (jednotky souboru), ale vždy v jiné hodnotě, a který se vyjadřuje pomocí obvyklých měřítek časových (rok, měsíc, den), je znak měřitelný, počet zaměstnanců v závodě živnostenském je znak sčítatelný.

Podle toho, kterého druhu jsou vyšetřené znaky, postupuje se při dalším popisu a rozboru odlišným způsobem.

## II. Rozložení statistických souborů.

Ze zákona velkých čísel vyplývá, že statistická šetření musí býti rozsáhlá, čímž docházíme k velkému počtu údajů, který je podmínkou spolehlivosti daného statistického materiálu a všech dedukcí, které z něho chceme získati. Abychom učinili tento rozsáhlý statistický materiál přehledným a mohli jej dále zpracovati, je jej třeba upravit v nějaký systém. To se děje tím, že seřazujeme soustavně statistické údaje podle nějakého hlediska, buď podle velikosti znaku od nejmenší hodnoty postupně až k největší nebo naopak, jde-li o soubor obsahující znak kvantitativní, nebo podle jiného hlediska, na př. místního, obsahuje-li soubor znak kvalitativní. Takto uspořádané statistické údaje se nazývají též statistickými řadami a sestavují se obyčejně ve formě statistických tabulek.

Věcné rozložení statistického souboru, obsahujícího znak kvantitativní, sestavuje se tak, že jednotlivé hodnoty znaku kvantitativního v souboru se vyskytující se seskupují ve skupiny zvané třídy a zjišťuje se četnost (frekvence) těchto skupin. Statistická tabulka má pak dva sloupce: v prvním jsou seřazeny podle velikosti různé hodnoty kvantitativního znaku, které byly pozorovány, obyčejně sdružené do tříd, v druhém sloupci je uveden počet jednotek, u nichž byla každá z těchto hodnot (hodnot třídy) při šetření zjištěna.

Je-li znak kvantitativní vyjádřen v celých jednotkách a nevykazuje-li příliš mnoho hodnot, dostáváme jednoduchý případ, kde třídy jsou dány přímo jednotkami znaku (tak zv. znak sčítatelný, měnící se přetržitě). V takovém případě, kde kvantitativní znak je dán přímo v celých jednotkách a rozpětí znaku (t. j. rozdíl mezi nejmenší a největší hodnotou znaku v souboru se vyskytující, t. zv. variační rozpětí) je malé, je též sám sebou dán interval třídy, t. j. rozdíl mezi třídami.

Častější případ však bývá ten, že kvantitativní znak se mění nepřetržitě (spojitě), t. j. nabývá hodnot velmi četných podle přesnosti šetření (je to znak měřitelný). V takovém případě je nejprve třeba zvoliti vhodný

třídní interval. Intervaly volíme zásadně stejné, aby bylo možno srovnávat mezi sebou četnosti tříd a postupovati k dalším statistickým operacím.

**Poznámka.** Ve výjimečných případech, kde to vyžadují zvláštní důvody, zásada rovnosti intervalů se nezachovává, pak ovšem nelze srovnávat četnosti tříd mezi sebou.

Kvantitativní znak měřitelný vyjadřujeme v obvyklých jednotkách míry, váhy, ceny a pod. a podle nich tvoříme též třídy. Na př. ceny třídíme tak, aby intervaly třídní byly 10 nebo 20 nebo 100 Kčs, pozorování anthropometrická tak, aby intervaly byly 2 nebo 5 nebo 10 centimetrů podle rozsahu souboru, věk obyvatelstva podle jednoho nebo 5 roků a pod.

Při sestavování tabulek rozložení četností postupujeme tak, že nejprve zvolíme interval třídy, stanovíme jeho velikost, dále určíme bod (hodnotu), od kterého vycházíme. Jako třídní interval bere se jednotka v případě, jestliže změny v ní vyjádřené jsou dosti značné. Jestliže však změny v kvantitativním znaku postupují příliš pomalu, je třeba zvoliti hodnotu jinou (na př. v uvedeném případě 5 let). V tomto případě se vyžaduje, aby hodnoty, spadající do každé třídy, příliš se neodchylovaly od středních hodnot tříd (na př. aritmetického průměru), dále aby interval byl dosti veliký, abychom se vyvarovali příliš velkého počtu tříd.

Tomuto požadavku vyhovuje obyčejně pravidlo, že rozdělíme rozdíl mezi nejvyšší a nejnižší hodnotou patnácti až dvacítipěti a tak dostaneme 15 až 25 tříd, které obyčejně postačí k popisu pozorovaného jevu, aniž se tabulka rozložení četností stala při tom nepřehlednou.

Pokud jde o stanovení bodu, od kterého při klasifikaci vycházíme, lze zvoliti číslo libovolné, obyčejně se však volí číslo celé. Důležitosti nabývá tato volba na př. tehdy, máme-li pochybnosti o správnosti statistického materiálu, abychom nekladli hranice tříd právě na hodnoty, které považujeme za nesprávné (na př. věk 40, 50, 60 let, ježto právě údaje o věku bývají podle zkušenosti nesprávně zaokrouhlovány na roky, končící desítkou).

Důležité jest, aby hraničné hodnoty mezi intervaly byly stanoveny přesně, t. j. tak, aby nevznikly pochybnosti, do které třídy má býti který pozorovaný případ zařaden. Aby bylo jasné vyznačeno, kde jest hraniční hodnota, označují se intervaly někdy způsobem:

0—

10—

20—

30— atd., což znamená, že veškeré hodnoty nižší než 10 zařazují se do třídy první, hodnoty počínající 10 a nižší než 20 do třídy druhé, hod-



noty počínající 20 a nižší než 30 do třídy třetí atd. Jindy se označují třídy konečnou hodnotou: —10

—20

—30 atd., což znamená, že veškeré hodnoty včetně 10 spadají do třídy první, veškeré hodnoty včetně 20 spadají do třídy druhé atd. Nebo se označují třídy střední hodnotou (průměrem intervalu) každé třídy, tedy ve stejném případě: 5

15

25

35 atd.

Počet zjištěných jednotek, vykazujících hodnoty v hranicích intervalu, dává nám četnost této třídy.

**Příklad č. 5. Věkové rozvrstvení mužů v Čechách podle pětiletých skupin věkových r. 1921.**

Dokončený rok věku	Ve věku označeném bylo přítomno	Ve věku označe- ném bylo z 1000 přítomných
0 — 4	217.492	67,8
5 — 9	315.461	98,4
10 — 14	368.495	114,9
15 — 19	354.167	110,4
20 — 24	304.307	94,9
25 — 29	246.483	76,8
30 — 34	219.020	68,3
35 — 39	206.940	64,5
40 — 44	196.464	61,3
45 — 49	182.504	56,9
50 — 54	160.548	50,1
55 — 59	136.147	42,4
60 — 64	114.885	35,8
65 — 69	83.083	25,9
70 — 74	54.324	16,9
75 — 79	29.880	9,3
80 — 84	12.097	3,8
85 — 89	2.912	0,9
90 — 94	621	0,2
95 — 99	96	0,0
100 a více	—	—
(neznámý)	(1.710)	(0,5)
	3,207.636	1000

**Příklad č. 6. Příklad nerovných intervalů:**

Obyvatelstvo v Čechách v r. 1930 podle velikosti obce pobytu.

Velikostní skupina obcí podle počtu obyvatelů	Obce	Přítomné obyvatelstvo	
		absolutně	ze 100 obyvatelů
1 — 100	168	14.198	0,2
101 — 200	1.510	236.236	3,3
201 — 300	1.745	432.288	6,1
301 — 500	2.123	825.244	11,6
501 — 1.000	1.721	1,189.525	16,7
1.001 — 2.000	709	963.651	13,6
2.001 — 5.000	336	1,031.400	14,5
5.001 — 10.000	79	554.685	7,8
10.001 — 20.000	35	472.972	6,7
20.001 — 50.000	14	425.650	6,0
50.001 — 100.000	—	—	—
100.001 a více ...	2	963.527	13,5
<b>Uhrn . . .</b>	<b>8.442</b>	<b>7,109.376</b>	<b>100,0</b>

Kdyby byly v tomto případě voleny intervaly stejné, vzniklo by buď příliš mnoho tříd, nebo by musily být zvoleny intervaly příliš velké, na př. o 2000 obyvatelích, a pak by velká většina všech případů připadla do třídy první a tím by se úplně zatemnilo rozložení obyvatelstva podle obcí.

Někdy se porušuje zásada stejných intervalů též proto, že se z původních tříd znaků vytvářejí samostatné útvary, ze kterých se stávají nové pojmy, na př. města s více než 100.000 obyvateli se nazývají velkoměsta, obce s méně než 2000 obyvateli se považují za obce venkovské, nebo zemědělské závody ve výměře do jednoho hektaru se označují jako závody trpasličí, zemědělské závody od 20 do 50 ha jako střední statky a pod., a tvoří pak samostatné pojmy kategorie sociologické, demografické nebo ekonomické.

Rozborem tabulek četností znaku kvantitativního můžeme sledovati pravidelnosti v rozložení četností. Zejména názorné je grafické vyjádření rozložení četností pomocí dvou os kolmých, z nichž na osu horizontální nanášíme stupně znaku kvantitativního (intervaly tříd), na kolmice vztyčené v bodech těchto intervalů pak nanášíme příslušné četnosti,

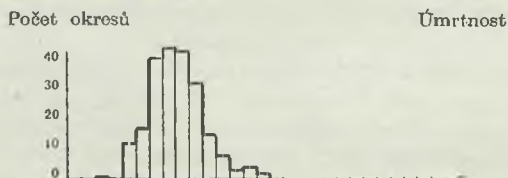
křivka, kterou se spojují body takto získané, nazývá se frekvenční křivkou. Jindy znázorňuje se rozložení četností polygonem četností čili histogramem, kde body na kolmicích vztyčených ve středu intervalů nespojují se křivkou, nýbrž krátkými vodorovnými přímkami (srv. kap. 13.).

Základním typem rozložení četností, zvaným též rozložení normální, je typ symetrického rozložení, u kterého četnosti postupně stoupají od nejnižších hodnot znaku ke středním hodnotám, kde četnost dosahuje vrcholu, a potom zase klesají s rostoucími hodnotami znaku. Rozložení toto se blíží theoretickému rozložení odchylek, jaké nám představuje Gaussova křivka. S úplnou symetrií nesetkáváme se však v souborech sociálních jevů snad nikdy, nýbrž pravidlem bývá typ více nebo méně asymetrický. Symetrickému rozložení vzdáleně se blíží rozložení četností soudních okresů v Čechách podle úmrtnosti v příkladu č. 7.

**Příklad č. 7. Roztřídění soudních okresů v Čechách podle úmrtnosti r. 1927** (Boháč, Studie o populaci, Praha 1928).

Počet zemřelých na 1000 obyvatel	Počet okresů s příslušnou úmrtností
méně než 11	1
11 — 12	10
12 — 13	18
13 — 14	42
14 — 15	43
15 — 16	43
16 — 17	32
17 — 18	15
18 — 19	8
19 — 20	3
20 — 21	4
21 — 22	2
22 — 23	1

V grafickém znázornění:





V souborech jevů sociálních se však častěji vyskytují případy, že četnosti jsou rozloženy nesouměrně tak, že při grafickém znázornění je vrchol křivky posunut na jednu nebo druhou stranu křivky, modus se při tom odchyluje od aritmetického průměru a mediánu (srv. kap. 7.). Jindy se vyskytují i frekvenční křivky nesouměrné o dvou nebo více vrcholech. Tento případ se vyskytuje obvykle u jevů, které nejsou stejnorodé, nýbrž se skládají z více souborů.

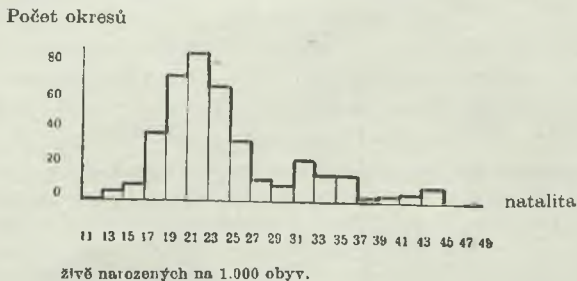
Konečně se vyskytují frekvenční křivky zcela nesouměrné, na př. takové, kde maximum četností (vrchol křivky) je položeno na jednom kraji křivky. Je to případ frekvenční křivky v podobě obráceného velkého J, která se vyskytuje často u souborů hospodářských (na př. rozložení četností daně důchodové podle výše zdaněných důchodů, rozložení zemědělských podniků podle velikosti výměry podniku a pod.). Největší četnost se vyskytuje v těchto případech mezi nejmenšími hodnotami souboru.

**Příklad č. 8. Roztřídění okresů v Čechách podle výše natality r. 1927** (Boháč l. c. str. 22).



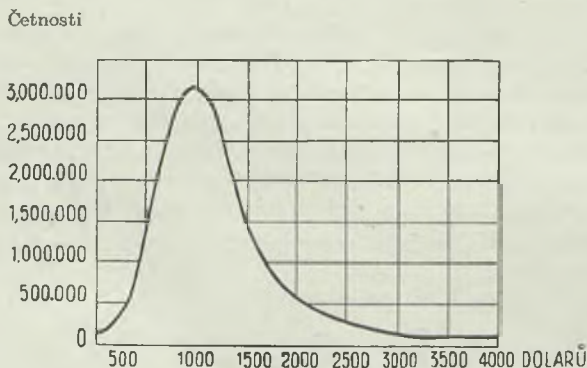
(příklad rozložení četností, jehož vrchol je posunut k jedné straně polygonu).

**Příklad č. 9. Roztřídění okresů v Československu podle výše natality r. 1926** (Boháč l. c. str. 22).



(příklad nesouměrného rozložení se dvěma vrcholy, které ukazují na to, že obyvatelstvo celého státu netvořilo po stránce demografické soubor zcela stejnorodý, zvláště se tu projevoval nepochybně vliv Zakarpatské Ukrajiny s odlišnými podmínkami demografickými).

**Příklad č. 10.** Rozložení četností vyjádřené frekvenční křivkou.



Rozložení četností poplatníků daně z příjmu ve Spojených státech r. 1918. Obsahuje všechny poplatníky, jejichž příjem byl nižší než 4000 dolarů. Interval třídy činí 500 dolarů. (*C. Mills*, *Statistical methods applied to economics and business*, New-York 1924, str. 86).

Statistický soubor se znakem kvantitativním může býti upraven též takovým způsobem, že se četnosti jednotlivých tříd stále k sobě přičítají čili kumulují se. V grafickém znázornění obdržíme pak zvláštní křivku četností, t. zv. křivku kumulační, která má podobu asi velkého S (Mills nazývá ji křivkou ogivní).

**Příklad č. 11.** *Křivka kumulační.* Rozložení četností souboru pil ve Spojených státech amerických r. 1920 podle výše pracovních nákladů (za 1000 stop zpracovaného dříví), dole polygon četností, nahoře kumulační křivka Mills l. c. str. 94).

**Poznámka.** Pro frekvenční křivky, které se blíží svým tvarem některému z typických tvarů frekvenčních křivek, lze nalézt analytické rovnice pomocí kterých je možno rozložení četností v daném souboru přesněji stanovit. Známá je na př. z historie národního hospodářství formule *Paretova*, která obsahuje analytickou rovnici pro rozložení četností typu J, charakteristické pro soubory hospo-

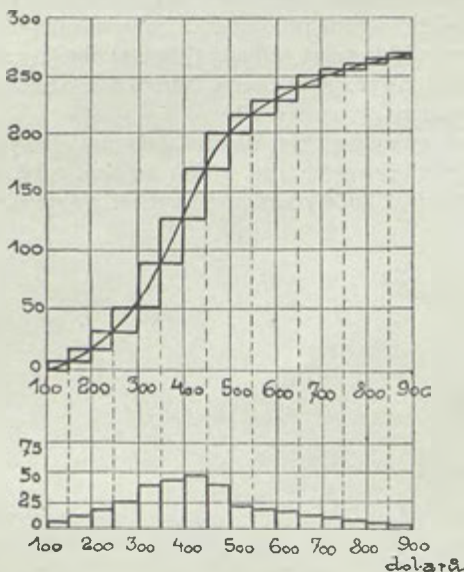
dářské. Tuto rovnici odvodil *Pareto* ze souboru důchodů a zní (je to křivka exponenciální):

$$y = Ax^{-a} \text{ nebo ve formě } \log y = \log A - a \log x$$

(*Pareto*, La courbe des revenus, Cours d'Economie politique 2, Lausanne 1897.)

Některé z těchto analytických rovnic mají význam též při sestavování úmrtních tabulek.

Příklad č. 11.



Statistický soubor lze někdy upravit též s *hlediska místního*. Vytvoří se t. zv. *řady místní*, ve kterých statistické hodnoty jsou uspořádány podle místních obvodů, na které se vztahují. Místní obvody jsou obvykle dány již samým administrativním rozdělením území, na kterém se statistické šetření konalo: územím obcí, okresů soudních, politických, zemí a jiných obvodů správních. Z těchto základních jednotek se někdy sestavují za účely lepšího přehledu vyšší jednotky místní, podle povahy souboru, na př. zmíněné přirozené oblasti pro soubory rázu zemědělského. V uspořádání řad místních rozhoduje do značné míry subjektivní hledisko statistikovo, uspořádá-li řadu podle nějakého směru geografického, či podle politického nebo hospodářského významu místních obvodů či jiného hlediska, nebo jen podle abecedního sledu jmen obcí, okresů, států.



V uspořádání takových místních řad může se tedy do značné míry uplatnit libovůle.

Obsahuje-li soubor znak *kvalitativní*, lze uspořádati jej podle obměn znaku kvalitativního. Obměny samy jsou ovšem uspořádány opět podle hlediska statistického, buď podle významu jednotlivých obměn, nebo podle zvláštního klasifikačního schematu, je-li počet obměn velmi velký, jako je tomu na př. u povolání obyvatelstva zjišťovaného při sčítání lidu, u druhů zboží zjišťovaných při statistice zahraničního obchodu. Každá hodnota řady vyjadřuje počet případů (četnost) obměny znaku, který se v souboru vyskytl (na př. počet Čechů, Němců a ostatních národností při sčítání lidu v souboru obyvatelstva).

Jsou-li soubory upraveny v *časovém sledu* statistických údajů podle obvyklých jednotek časových, dní, týdnů, měsíců, roků, mluvíme o *řadách časových*, o nichž bude podrobnější výklad v kapitole 9.

## STATISTICKÉ CHARAKTERISTIKY.

Jako statistické charakteristiky označujeme hodnoty, které vyjadřují a charakterisují soubor. Vypočítáváme je, abychom zjednodušili soubor složený z mnoha statistických údajů (hodnot) a zároveň jej charakterisovali vhodnou hodnotou, která nivelisuje odchylky v souboru se vyskytující.

### I. Charakteristiky polohy.

1. *Aritmetický průměr.* Nejjednodušší a nejčastěji používanou statistickou charakteristikou je aritmetický průměr. Aritmetický průměr obdržíme, dělíme-li součet všech hodnot souboru počtem hodnot, tedy označíme-li jednotlivé hodnoty statistického souboru písmeny  $x_1, x_2, x_3, x_4$  až  $x_r$ , je vzorec aritmetického průměru

$$A = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_r}{r} \quad \text{nebo v podobě } A = \frac{\sum (X)}{r},$$

kde písmenem  $\sum$  označujeme součet těch hodnot, které jsou za ním uvedeny v závorce. (*Místo písmene  $A$  užívá se pro aritmetický průměr hodnot  $X_1, X_2$  atd. též označení  $\bar{X}$ .*)

Nejdůležitější vlastnosti aritmetického průměru jsou tyto: úhrn veškerých odchylek (positivních i negativních) jednotlivých hodnot souboru od aritmetického průměru se rovná nule; součet dvojmočí těchto odchylek je menší než součet dvojmočí odchylek jednotlivých hodnot souboru, vypočtených od jakékoli jiné hodnoty souboru (čili je minimum). Říkáme též, že aritmetický průměr je z celého souboru hodnota nejpravděpodobnější.

Obsahuje-li soubor znak kvantitativní, je třeba jednotlivé hodnoty znaku násobiti příslušnou jejich četností, jmenovatelem je pak ovšem součet všech četností znaku v souboru. Takový aritmetický průměr můžeme

vyjádřiti vzorcem  $A = \frac{\sum (fx)}{r}$ , kde  $f$  znamená četnost každé hodnoty  $x$ .

Aritmetický průměr, při kterém bereme zřetel k četnostem hodnot, nazývá se průměr vážený (z anglického weighted average), poněvadž v něm dáváme jednotlivým hodnotám tu váhu (význam), jakou četností jsou zastoupeny v celém souboru. Volný výběr vážení či nevážení průměry máme zvláště ve statistice cenové.

U znaků kvantitativních, měnících se spojitě (nepřetržitě), často i u znaků měnících se přetržitě, avšak u nichž je rozpětí hodnot veliké, naráží vypočítávání aritmetického průměru na obtíže. Uvnitř každého intervalu jest totiž řada hodnot, jejichž velikost ani četnosti neznáme. V takových případech pomáháme si často tím, že předpokládáme, že hodnoty uvnitř intervalu rostou pravidelně, a že tudíž aritmetický průměr hodnot znaku a jejich četností se rovná aritmetickému průměru horní a dolní hranice intervalu (tento předpoklad nemusí se ovšem shodovati se skutečností, čímž vzniká jistá nepřesnost takto vypočtené hodnoty).

Za základ výpočtu aritmetického průměru v tomto případě bereme střed intervalu, který násobíme četností každé třídy. Součet těchto součinů pak dělíme počtem všech případů (četností).

**Příklad č. 12.** *Počet vdaných žen v Čechách podle skupin věku r. 1930.*  
(Částečný soubor.)

Věk v dokončených letech	Počet vdaných žen
19 a méně	8.629
20 až 24	118.390
25 až 29	221.586
30 až 34	240.283
35 až 39	211.308
40 až 44	179.473
45 až 49	158.358
<hr/>	
Úhrn	1,138.027

Středů intervalů každé skupiny násobíme četností skupiny (třídy):

střed intervalu skupiny	četnost skupiny
17	8.629
22	118.390
27	221.586
32	240.283
37	211.308
42	179.473
47	158.358
<hr/>	
Úhrn	1,138.027



Aritmetický průměr se rovná

$$A = \frac{17 \times 8.629 + 22 \times 118.390 + 27 \times 221.586 + 32 \times 240.283 + 37 \times 211.308 + 42 \times 179.473 + 47 \times 158.358}{1,138.027} = 34,466 \text{ let.}$$

Výpočet aritmetického průměru můžeme si zjednodušiti pomocí prozatímního aritmetického průměru, t. j. vhodně zvoleného počátku, čímž se vyhneme obtížnému sčítání a dělení velkými čísly. K tomu slouží vzorec, zakládající se na uvedených vlastnostech aritmetického průměru, že úhrn odchylek od něho se rovná nule:

$$A = \frac{(x_1 - A_0) + (x_2 - A_0) + (x_3 - A_0) + \dots + (x_r - A_0)}{r} + A_0,$$

t. j. libovolnou hodnotu souboru bez výpočtu zvolíme za prozatímní průměr a označíme  $A_0$ , ale tak, aby přibližně, podle odhadu, nebyla příliš vzdálena od skutečného aritmetického průměru. Tuto hodnotu odečteme od každého člena statistické řady a pak ji k celkovému průměru těchto rozdílů opět přičteme. Výhoda tohoto postupu je patrná, že se značně zmenšují čísla, se kterými počítáme a celý počet se ulehčuje.

Podobným způsobem lze postupovati též tehdy, je-li soubor rozdělen ve třídy podle znaku kvantitativního. Rozdíly od prozatímního průměru počítáme pak v jednotkách třídy a výsledek násobíme intervalem třídy.

Ve výše uvedeném příkladě bychom postupovali tímto způsobem:

Střed intervalu třídy	Rozdíl od prozatímního průměru $A_0$	Četnost třídy $f$	Součin $a f$
17	-3	8.629	- 25.887
22	-2	118.390	-236.780
27	-1	221.586	-221.586
32	0	240.283	-484.253
			úhrn záporných
37	+1	221.308	+ 211.308
42	+2	179.473	+ 358.946
47	+3	158.358	+ 475.074
			+1,045.328
			úhrn kladných
			- 484.253
			+ 1,045.328
			úhrn + 561.075

Za prozatímní průměr zvolíme v tomto případě střed třídy  $A = 32$ , od tohoto průměru počítáme rozdíly tříd (sloupec druhý). Rozdíly pak násobíme četnostmi (sloupec čtvrtý). Součiny sečteme, záporné i kladné zvlášť, konečný součet dělíme pak součtem všech četností:

$$561.075 : 1,138.027 = 0,493 \text{ jednotek třídních.}$$

Výsledek násobíme třídním intervalem  $I = 5$

$$0,493 \times 5 = 2,465$$

a přičteme nyní ke zvolenému prozatímnímu průměru  $A_0$

$$A = 32 + 2,465 = 34,465$$

(malý rozdíl ve třetím desetinném místě je způsoben nepřesností krácení při dělení).

Aritmetický průměr celkového souboru, skládajícího se ze souborů dílčích, rovná se součtu aritmetických průměrů těchto dílčích souborů:

$$A = A_1 + A_2 + A_3.$$

Aritmetický průměr jako veličina vypočtená je hodnota abstraktní, to znamená, že se ve skutečnosti nemusí taková hodnota v celém souboru vyskytnouti, čímž trpí aritmetický průměr tím více, čím je rozložen souboru nesouměrněji.

Jiné statistické charakteristiky abstraktní, jichž se ve statistice častěji používá, jsou zvlášť geometrický průměr, dále harmonický a kvadratický průměr.

2. Průměr geometrický roven je součinu všech členů souboru, od mocněnému odmocnitelem, který se rovná počtu těchto členů podle vzorce

$$G = \sqrt[r]{x_1 \times x_2 \times x_3 \times x_4 \times \dots \times x_r}.$$

Geometrický průměr má hlavně ty vlastnosti statisticky významné, že není tak citlivý k extrémním hodnotám a že je nižší než průměr aritmetický z téhož souboru. Používá se ho zvlášť ve statistice cenové a k se-strojování indexů.

Obsahuje-li soubor znak kvantitativní a je-li rozdělen na třídy, vypočítává se geometrický průměr podle vzorce

$$G = \sqrt[r]{x_1' \times x_2' \times x_3' \times \dots \times x_r'}$$

**Poznámka.** Zřídka se používá průměru harmonického, který je vlastně zvrátnou (reciproční) hodnotou průměru aritmetického podle vzorce

$$H = \frac{r}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_r}}$$

a průměru kvadratického, rovněž v podstatě průměru aritmetického, jehož jednotliví členové se umocňují dvěma a celek se odmocňuje dvěma podle vzorce

$$AQ = \sqrt[n]{x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2}$$

**Příklad č. 13.** Jednoduchý příklad vypočítávání průměrů.

Statistická řada skládá se z hodnot: 34, 38, 42, 46.

$$\text{Aritmetický průměr } A = \frac{34 + 38 + 42 + 46}{4} = 40$$

Průměr geometrický počítá se nejvýhodněji v logaritmech:

$$\log G = \frac{1,531 + 1,579 + 1,623 + 1,662}{4} = 1,599$$

$$G = 39,75.$$

Průměr harmonický týchž čísel se rovná

$$H = \frac{4}{0,029 + 0,026 + 0,023 + 0,021} = \frac{4}{0,099} = 40,04 \text{ (zkráceným dělením)}$$

Průměr kvadratický týchž čísel se rovná

$$AQ = \sqrt{\frac{1156 + 1444 + 1764 + 2116}{4}} = \sqrt{\frac{6480}{4}} = 40,249$$

Pozorujeme, že průměry vypočítávané různým způsobem při správném výpočtu příliš se od sebe neliší, je-li rozložení souboru symetrické.

3. Jinou statistickou charakteristikou je medián, hodnota prostřední. Je to hodnota ležící uprostřed, seřadíme-li všechny členy souboru podle velikosti znaku kvantitativního, jinak definováno, hodnota, vedle které nalezneme v souboru stejné množství hodnot jak vyšších, tak nižších.

Medián zjistíme snadno, je-li počet členů statistické řady lichý; je-li sudý, nahrazujeme jej aritmetickým průměrem dvou prostředních členů.

Používání mediánu má proti aritmetickému průměru některé výhody, z nichž možno jako nejvýznamnější uvést tyto:

1. Snadný a rychlý způsob, jakým jej možno zjistiti bez dlouhého počítání.

2. Výhoda, že není případně třeba znáti krajní členy statistické řady, postačí, známe-li pouze počet členů, ovšem za předpokladu, že jsou i neznámí členové uspořádání podle velikosti.

3. Malá citlivost, případně necitlivost vůči krajním hodnotám značně se odchylujícím od hodnot ostatních.

Naproti tomu má medián proti aritmetickému průměru některé nevýhody:

1. Násobíme-li medián počtem případů, nedostaneme úplnou řadu statistickou, takže se medián nehodí pro srovnání dvou statistických řad.



2. Necitlivost mediánu vůči krajním hodnotám může být v některých případech opět nevýhodou.

3. Soubor musí být dříve uspořádan, údaje musí být nejprve srovnány podle velikosti.

4. Medián je nepřesný v případě, jestliže se v řadě vyskytuje více stejně velkých hodnot.

Jestliže máme soubor rozdělený ve třídy, je sice snadno zjistit třídu, do které medián spadá, ale je pak ještě třeba zjistit uvnitř třídy hodnotu, která je skutečným mediánem. Výpočet této hodnoty se děje podle vzorce:

$$Me = l \times \frac{b + c - a}{2b},$$
 kde  $l$  znamená opět rozpětí intervalu,  $b$  četnost třídy, do které medián spadá,  $c$  četnosti všech tříd následujících této třídě,  $a$  četnosti všech tříd předcházejících této třídě.

Poznámka. Medián z hodnot  $x_1, x_2, x_3 \dots x_n$  označuje se též symbolem  $\tilde{x}$ .

Vedle mediánu používá se někdy ještě podřadnějších hodnot, které rozdělují statistickou řadu na více částí, zvláště kvartilů, decilů atd.

*Kvartily* nazýváme ony hodnoty, které rozdělují statistickou řadu (usměrněnou) na čtyři stejné díly; při tom druhý kvartil spadá v jedno s mediánem. Kvartily vypočítávají se podle vzorce

$$Q_1 = \frac{n+1}{4} \text{ (první kvartil), } Q_3 = \frac{3(n+1)}{4} \text{ (třetí kvartil), medián}$$

$$Me = \frac{2(n+1)}{4} \text{ (současně druhý kvartil).}$$

Analogického vzorce jako pro výpočet mediánu bylo by třeba použít pro výpočet kvartilů:

$$\text{pro první kvartil } Q_1 = l \times \frac{b + c - 3a}{4b}$$

$$\text{a pro třetí kvartil } Q_3 = l \times \frac{3(b + c) - a}{4b}$$

4. Jiná střední hodnota, často používaná pro soubory obsahující znak kvantitativní, je *modus* čili *nejčetnější hodnota* znaku (někdy též zv. *dominanta*). Všeobecně lze ji označit též jako hodnotu nejčastější, neboť podle této střední hodnoty se obvykle statistický soubor posuzuje. Mluvíme-li o průměrné nebo nejčastější mzdě dělníků nějaké továrny, máme na mysli právě modus, t. j. mzdu, která skutečně největšímu počtu dělníků je vyplácena, nikoli aritmetický průměr mezd. Podobně průměrný věk žáků nějaké třídy, průměrná velikost vojnů atd. posuzuje se obvykle podle modu.

Modus má podobné výhody jako medián, zejména je necitlivý vůči málo četným krajním hodnotám, dále lze ho použít i tehdy, jestliže neznáme četnosti všech členů statistického souboru, stačí, víme-li, že ony četnosti jsou menší než četnosti, které známe.

Nevýhodou modu je nesnadnost jeho přesného určení v případě, že znak roste nepřetržitě. Pro výpočet modu existuje řada method, které vesměs vycházejí z určitých předpokladů. Nejjednodušší metoda se zakládá na předpokladu, že četnosti uvnitř třídy jsou rozloženy rovnoměrně, potom postačí vzít za modus střed intervalu třídy s největší četností.

Jiná metoda záleží v tom, že k dolní hranici třídy s největší četností připočteme diferenci  $d$  podle vzorce

$$d = - \frac{(b - a)}{(c - b) - (b - a)},$$

kde  $b$  značí četnost třídy nejvíce zastoupené,  $a$  četnost třídy předcházející a  $c$  četnost třídy následující. Tuto diferenci násobíme pak opět rozpětím intervalu.

Také hodnota modu se mění podle toho, jaký interval při původní úpravě souboru zvolíme, takže není zcela jednoznačná.

#### Příklad č. 14. Výpočet mediánu a modu.

*Pronajaté byty o 1 pokoji s kuchyní v Praze r. 1927 (pro něž bylo stavební povolení uděleno před 28. 1. 1917) podle výše činže.*

Roční činžovní hodnota činila v Kčs	Počet pronajatých bytů
do 200	2.006
přes 200 do 400	12.713
přes 400 do 600	20.712
přes 600 do 800	12.228
přes 800 do 1000	4.931
přes 1000 do 1200	1.849
přes 1200	1.759

(Upravená tabulka podle Konopáče, Nájemné ve Velké Praze v letech 1923 až 1930, Praha 1931, str. 17.)

Z této tabulky rozložení četností vidíme, že třída třetí obsahuje v sobě jak medián, tak modus. Předpokládáme-li, že počet bytů uvnitř této třídy roste celkem pravidelně s velikostí činže, k čemuž nás opravňuje okolnost, že druhá i čtvrtá třída mají téměř stejnou četnost, zjistíme přibližnou hodnotu modu aritmetickým průměrem dolní a horní meze intervalu třídy třetí, což dává nám hodnotu 500 Kčs.

Jestliže použijeme pro výpočet modu vzorce výše uvedeného, dostaneme

$$d = - \frac{20.712 - 12.713}{(12.228 - 20.712) - (26.712 - 12.713)} = - \frac{7.999}{- 16.483} = 0,485;$$

tuto diferenci násobíme intervalem třídním čili 200.

Modus se pak rovná dolní hranici třídy s největší četností zvětšené o takto vypočtenou hodnotu:

$$M_0 = 400 + (0,485 \times 200) = 497 \text{ Kčs.}$$

**Poznámka.** Modus z hodnot  $x_1, x_2, x_3, \dots, x_r$  označuje se též  $x$ .

Medián uvnitř třídy podle vzorce uvedeného na str. 56 se rovná

$$M\acute{e} = I \times \frac{b + c - a}{2b}$$

$$M\acute{e} = 200 \times \frac{20.712 + 20.767 - 14.719}{2 \times 20.712} = 129.$$

Vypočtenou hodnotu přičteme opět k dolní hranici třídy, ve které je medián obsažen a dostaneme

$$M\acute{e} = 400 + 129 = 529 \text{ Kčs.}$$

Je-li statistický soubor zcela symetrický, spadají aritmetický průměr, modus i medián v jedno...

## II. Charakteristiky rozptylu.

Rozptylem (dispersí) nazývá se rozložení hodnot souboru v mezích daných nejmenší a největší hodnotou (v mezích variačního rozpětí). Měrami rozptylu zkoumáme rozložení jednotlivých hodnot znaku kolem hodnot středních. Čím je rozptyl menší, tím jsou hodnoty souboru úže seskupeny kolem charakteristiky souboru.

### 1. Průměrná odchylka.

Průměrná odchylka je aritmetický průměr odchylek jednotlivých hodnot souboru od některé charakteristiky polohy, nejčastěji mediánu. Odchylky zaznamenáváme v jejich absolutní hodnotě, neboť jinak by se odchylky pozitivní a negativní vzájemně zrušily (vyrovnaly).

$$\vartheta = \frac{(x_1 - M\acute{e}) + (x_2 - M\acute{e}) + \dots + (x_r - M\acute{e})}{r}$$

jsou-li dány četnosti hodnot znaku  $f_1, f_2, f_3, f_4, f_5, \dots, f_n$  dohromady =  $r$ ,

$$\vartheta = \frac{(x_1 - M\acute{e}) f_1 + (x_2 - M\acute{e}) f_2 + \dots + (x_r - M\acute{e}) f_n}{r}$$



## 2. Směrodatná (standardní) odchylka.

Obyčejně se však používá k měření rozptylu směrodatné (standardní) odchylky, která se všeobecně označuje řeckým písmenem malým sigma  $\sigma$ .

Směrodatná odchylka je sestrojena v podstatě na stejném základě jako průměrná odchylka, avšak místo jednoduchých odchylek hodnot bereme za základ čtverce odchylek, které počítáme od aritmetického průměru souboru, celý průměr se pak odmocňuje dvěma.

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \dots + (x_r - \bar{x})^2}{r}} \quad \text{nebo} \quad \sqrt{\frac{\sum (x - \bar{x})^2}{r}}$$

kde  $x_1, x_2, x_3, x_r$  atd. značí opět členy souboru a  $\bar{x}$  aritmetický průměr jeho,  $r$  celkový počet hodnot souboru.

Jsou-li dány četnosti hodnot statistického souboru, je třeba násobiti čtverce odchylek příslušnými četnostmi podle vzorce

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + (x_3 - \bar{x})^2 f_3 + \dots + (x_r - \bar{x})^2 f_r}{r}}$$

Je-li soubor rozdělen podle kvantitativního znaku rostoucího nepřetržitě na třídy, je výpočet směrodatné odchylky ztížen tím, že musíme vycházeti od středů tříd, podobně jako tomu bylo při výpočtu aritmetického průměru, tedy od předpokladu, že střední hodnota každého intervalu je průměrem z hodnot znaku v každé třídě. Dále bychom dostali často velmi velká čísla pro jednotlivé odchylky, kdybychom je počítali od skutečného aritmetického průměru pro všechny střední třídy. Proto se používá zkráceného výpočtu pomocí prozatímního aritmetického průměru, podobně jako při výpočtu tohoto průměru.

Za základ výpočtu vezme se střed intervalu, do kterého spadá aritmetický průměr, za pomocný aritmetický průměr, a od něho se počítají odchylky. Od takto vypočtené směrodatné odchylky je třeba ještě odečísti rozdíl mezi prozatímním průměrem a skutečným aritmetickým průměrem souboru, vyjádřený v jednotkách intervalu. Poněvadž se celý výpočet děje v jednotkách intervalu, je třeba výsledek násobiti ještě rozptím intervalu.

*Schematický příklad výpočtu průměrné odchylky.*

**Příklad č. 15.** Máme statistický soubor, skládající se ze členů

19, 25, 26, 28, 32, 33, 40.

Aritmetický průměr  $A = 29$ .

Odchylky jednotlivých členů od aritmetického průměru jsou

19	$\xi_1 = 10$
25	$\xi_2 = 4$
26	$\xi_3 = 3$
28	$\xi_4 = 1$
32	$\xi_5 = 3$
33	$\xi_6 = 4$
40	$\xi_7 = 11$
<hr/>	
	$\Sigma (\xi) = 36$

$$\vartheta = \frac{36}{7} = 5,142.$$

**Příklad č. 16.** Příklad výpočtu směrodatné odchylky souboru rozděleného ve třídy pomocí prozatímního průměru (srv. př. č. 12 na str. 52).

*Počet vdaných žen v Čechách podle skupin věku r. 1930.*

Střed intervalu věkové třídy	Rozdíl od aritmetického zat. průměru $\xi$	Dvojnásobek rozdílu $\xi^2$	Četnost třídy $f$	Součin $\xi^2 \cdot f$
17	-3	9	8.629	77.661
22	-2	4	118.390	473.560
27	-1	1	221.586	221.586
32	0	0	240.283	0
37	+1	1	211.308	211.308
42	+2	4	179.473	717.892
47	+3	9	158.358	1,425.222

$$\Sigma(f) = r = 1,138.027 \quad \Sigma(\xi^2 \cdot f) = 3,127.229$$

$$\frac{\Sigma(\xi^2 \cdot f)}{r} = \frac{3,127.229}{1,138.027} = 2,748$$

$$\sigma^2 = \frac{\Sigma(\xi^2 \cdot f)}{r} - C_A^2$$

kde  $C_A$  značí rozdíl mezi prozatímním aritmetickým průměrem a skutečným aritmetickým průměrem. Tento rozdíl se v daném příkladě (v. str. 54) rovná  $C_A = 0,493$  jednotek třídních

$$\sigma^2 = 2,748 - 0,243 = 2,505$$

$$\sigma = \sqrt{2,505} = 1,5827.$$

Tento výpočet směrodatné odchylky je značně snazší, než kdybychom počítali rozdíly středů intervalů jednotlivých tříd od skutečného aritmetického průměru, neboť pracujeme s podstatně nižšími čísly při umocňování, násobení i dělení.

### 3. Kvartilní odchylka.

Rozptyl a jeho souměrnost lze jednoduchým způsobem vyjádřit také pomocí vzdálenosti obou kvartilů. Označíme-li kvartily opět písmeny  $Q_1$  pro první kvartil a  $Q_3$  pro třetí kvartil, rovná se kvartilní odchylka polovičnímu rozdílu obou kvartilů

$$I_Q = \frac{Q_3 - Q_1}{2}.$$

### 4. Koeficient rozptylu (variační koeficient).

Kombinací směrodatné odchylky s aritmetickým průměrem je poměrná míra rozptylu, nazývaná koeficientem rozptylu či variačním koeficientem:

$$V = \frac{\sigma}{A} \times 100.$$

Tohoto koeficientu se užívá s výhodou ke srovnání rozptylů dvou nebo více souborů, neboť vyjadřuje rozptyl v procentech aritmetického průměru.

### 5. Míry šikmosti rozptylu.

Rozptyl statistického souboru může býti ještě charakterisován stupněm souměrnosti či nesouměrnosti rozložení okolo hodnot středních. Nesouměrnost, šikmost či křivost sklonu křivky rozložení četností (anglický termín pro šikmost rozptylu, často užívaný, je *skewness*) vyjadřuje se nejčastěji srovnáním aritmetického průměru a modu se směrodatnou

odchylkou souboru:  $\frac{A - M_0}{\sigma}$ , t. j. rozdíl mezi aritmetickým průměrem

a modem se uvádí v poměr ke směrodatné odchylce. Při úplné souměrnosti křivky rozložení spadají aritmetický průměr a modus ovšem v jedno, a míra šikmosti se pak rovná nule.

Literatura ke kap. 6—7. Vedle spisů v textu citovaných Janko: Jak vytváří statistika obrazy světa.



## Kapitola 8.

# ČÍSLA POMĚRNÁ A ČÍSLA INDEXNÍ.

### 1. Čísla poměrná.

Vedle čísel absolutních, která vznikají přímo jako výsledek statistických šetření a jsou elementární podobou statistických údajů, používá se k popisu a rozboru statistických souborů čísel poměrných čili relativních. Čísla poměrnými nazýváme čísla, která vyjadřují vzájemné vztahy statistických údajů. Poměrnými čísla vyjadřujeme tedy buď vztahy mezi částí souboru a jeho celkem, nebo mezi různými částmi téhož souboru, nebo mezi tímž souborem v různých časových obdobích, nebo konečně mezi různými soubory, mezi kterými předpokládáme nějakou logickou spojitost.

Vztah mezi částí souboru, vykazující určitý znak, a mezi jeho celkem, takže v čitateli je obsažena část souboru, ve jmenovateli celý soubor, nazývá se poměrnou četností znaku (jindy poměrnými četnostmi analytickými) a představuje vyjádření poměru empirické pravděpodobnosti. Na př. poměr mužů, t. j. jedinců, vykazujících znak mužského pohlaví k počtu všech jedinců, t. j. celému souboru obyvatelstva.

Poměrná čísla mohou vyjadřovati též vztah jedné části souboru, vykazující určitý znak, k jiné části souboru, vykazující jiný znak, na př. poměr novorozených chlapců k novorozeným děvčatům (jsou to dvě části téhož souboru novorozených dětí).<sup>1)</sup>

Poměrná čísla mohou vyjadřovati též vztah mezi tímž souborem v různých obdobích časových, což je jeden a to nejčastější z případů in-

---

<sup>1)</sup> *Kohn* ukazuje, jak tato poměrná čísla mohou býti snadno převedena na poměrné četnosti, l. c. str. 16.

dexních čísel, jimž bude věnována druhá část této kapitoly. V tomto případě se mluví také o vztazích koordinovaných souborů.

Poměrná čísla mohou dále vyjadřovati vztahy mezi různými soubory, které spolu mají nějakou souvislost, na př. vztah mezi počtem obyvatelstva a rozlohou státního území, na kterém toto obyvatelstvo je usazeno. Počet obyvatelstva připadající na jeden čtverečný kilometr označuje se obyčejně jako hustota obyvatelstva (v Čechách činil r. 1930 137, na Moravě a ve Slezsku 133, na Slovensku 68). Jiný příklad: počet sňatků nebo úmrtí uvedený v poměr k počtu obyvatelstva. Taková poměrná čísla nazývají se sňatečností nebo úmrtností, nejsou však tak přesná jako poměrné četnosti, poněvadž v čitateli je soubor událostí zachycený během celého roku, kdežto ve jmenovateli soubor jevů trvalých, zjištěných k počátku roku nebo k jinému datu (obyvatelstvo se během roku mění porody a úmrtími i migrací).

Stejnorodé statistické soubory lze spojití pomocí společného jmenovatele. Ve statistice se nazývá společným jmenovatelem společná jednotka míry pro rozmanité veličiny, na př. velmi různé druhy zboží v zahraničním obchodě se převádějí na společného jmenovatele pomocí jednotky měny, neboť váha různého zboží dohromady sečítaného nedá žádného obrazu o zahraničním obchodě, nejvýše o dopravním výkonu. potřebném pro tento obchod.

Jako speciální poměrná čísla se označují poměrná čísla všech uvedených druhů, ve kterých se část souboru nebo soubor jiného druhu uvádí v poměr k dílčímu souboru, ke kterému má věcný vztah, na př. poměrné číslo počtu narozených dětí k dílčímu souboru obyvatelstva, omezenému na ženy ve věku plodném (obyčejně ve věku od 15—45 let). Toto speciální poměrné číslo se pak označuje jako plodnost proti natalitě (nebo porodnosti), které je poměrným číslem obecným (poměr počtu narozených dětí ke všemu obyvatelstvu). Takový vztah může býti ještě užším vymezováním souboru ještě dále zužován, na př. přidáváním znaku žen vdaných, čímž dostáváme ještě užší (a ještě více homogenní) soubor dílčí.

Poměrná čísla se vyjadřují buď desetinným zlomkem nebo v procentech nebo promile, podle toho, položí-li se soubor, ke kterému uvádíme jiný soubor ve vztah, rovný 1 nebo 100 nebo 1000. Chceme-li tedy zjistiti, kolik procent z čísla  $X$  činí číslo  $Y$ , násobíme číslo  $Y$  stem a součín dělíme číslem  $X$ , zlomek  $\frac{100 Y}{X}$  dává nám hledané procento. Podobně zjistíme promile obdobným způsobem: násobíme-li číslo  $Y$  tisícem a dělíme číslem  $X$ .



## 2. Čísla indexní.

Zvláštním případem poměrných čísel jsou čísla indexní. Indexní čísla (ukazatelé, z angl. index numbers) jsou čísla, která svými změnami a svým pohybem vyjadřují změny souboru jevů, ke kterému jsou v určitém vztahu. Jindy se definují indexní čísla jako čísla, která svými změnami charakterisují vzrůst nebo pokles nějaké veličiny (souboru), která se sama vymyká přesnému měření. Indexní čísla slouží ke srovnání stejnorodých statistických hodnot, a to nejčastěji ke srovnání časovému, někdy též místnímu nebo věcnému. Pravým účelem indexních čísel je tedy srovnání.

Pojem indexních čísel se rozšiřuje často i na poměrná čísla, která vyjadřují změny v jediné statistické řadě v poměru ke zvolené hodnotě základní, a taková indexní čísla se nazývají jednoduchými indexními čísly, na př. cena pšenice podle záznamů plodinové bursy se srovnává každý týden s cenou pšenice, zaznamenanou dne 1. ledna 1930 tím způsobem, že se vyjadřuje, kolik procent z ceny pšenice dne 1. ledna 1930 činí cena pšenice v každém časovém období. Nebo množství uhlí vytěženého v každém roce se uvádí stejným způsobem v poměr ke množství vytěženému v roce, který jsme zvolili za rok základní. Takový jednoduchý index se skládá z řady prostých koordinovaných čísel poměrných, vyjadřujících poměr ceny pšenice nebo množství vytěženého uhlí ke členu základnímu.

Příkladem indexních čísel ve smyslu uvedených výše definic (nazývají se též úhrnná indexní čísla) jsou indexy cenové, kde z průměrů změn v cenách jednotlivých statků snažíme se vyšetřiti změny v jiném souboru, totiž cenové hladině představující kupní sílu peněz nebo životní náklady.

Obvykle se násobí základní hodnota nebo průměr stem nebo tisícem, čímž dostaneme stejně vyjádřená indexní čísla.

Při vypočítávání jednoduchých indexních čísel platí to, co bylo uvedeno o poměrných číslech. Všeobecně je však třeba pečlivě vybrati základ indexů, období nějak zvláště významné a důležité, není-li takového, bere se průměr několika období, abychem se vyhnuli možným nahodilým odchylkám. Často řídí se ovšem volba základního období povahou statistického pozorování nebo jeho účelem, na př. při indexech cenových bere se dosud často za základ poslední mírový rok před první světovou válkou se zřetelem na pronikavé změny cenové, které od té doby nastaly. Je-li řada uzavřená, konečná, možno vzíti za základ indexních čísel též průměr všech členů řady.



Pokud jde o indexy úhrnné, sestavují se tak, že vypočítává se průměr z jednotlivých poměrných čísel, vyjadřujících změny v dílčích soubořech, na př. u československého velkoobchodního indexu se vypočítával aritmetický (a též geometrický) průměr ze 69 poměrných čísel, vyjadřujících poměr cen 69 druhů zboží v určitém měsíci k cenám týchž druhů zboží dne 1. července 1914.

Řetězovými indexy se nazývají taková indexní čísla, kde každé číslo se uvádí v poměr k předcházejícímu, takže základna srovnání se mění od čísla k číslu (indexní čísla o pohyblivé základně).

Největšího rozšíření došla indexní čísla ve statistice cenové, neboť právě kupní síla peněz a hladinu nákladů životních lze měřiti jen pomocí jich. Toto rozsáhlé používání indexních čísel v cenové statistice vyvolalo v theorii i praxi velkou řadu vzorců pro vypočítávání indexních čísel, které se zakládají na různých kombinacích poměrných čísel a středních hodnot. V praxi převládlo však používání aritmetického nebo geometrického průměru neváženého nebo váženého.

Také kritérii správnosti vzorců indexních čísel, zejména zkoušky se záměnou času a záměnou činitelů, které sestavil *Irving Fisher*,<sup>1)</sup> se v praxi nevalně dbá.

#### Příklad č. 17. Příklad jednoduchého indexu:

*Těžba kamenného uhlí v letech 1926—1935.*

##### *Index těžby.*

Rok		(denní průměr 1929 = 100)
1926	14,177.134 tun	86,0
1927	14,016.330 „	87,0
1928	14,567.898 „	90,0
1929	16,548.227 „	100,0
1930	14,468.519 „	86,4
1931	13,165.051 „	78,8
1932	11,032.172 „	65,3
1933	10,627.357 „	63,5
1934	10,788.880 „	64,1
1935	10,894.483 „	65,5

<sup>1)</sup> *I. Fisher*, *The making of Index Numbers*, N. York 1922, v české literatuře podrobněji o kritériích správnosti cenových indexů a t. zv. zkoušce interkalace *Kohn*, l. c. str. 114 sl.

**Příklad č. 18. Příklad úhrnných indexních čísel:**

*Index velkoobchodních cen v jednotlivých měsících r. 1945.*

(ceny jsou zjišťovány vždy k 1. dni každého měsíce, základem jsou ceny, zjištěné k 1. červenci 1914 = 100, vypočteny na základě geometrického průměru).

Měsíc:	
Leden	1091
Únor	1089
Březen	1089
Duben	1086
Květen	1087
Červen	1092
Červenec	1105
Srpen	1102
Září	1111
Říjen	1132
Listopad	1150
Prosinec	1488
Celoroční průměr 1945	1214

**Literatura ke kapitole 8.** Vedle spisů v textu citovaných *Flaskämper* Theorie der Indexzahlen, Berlin 1928, týž Statistik, *Meyer's* Wörterbücher, Lipsko 1930, *Winkler*, Die statistischen Verhältniszahlen, Lipsko 1928, *Wagemann* Der Narrenspiegel der Statistik, Hamburk 1935, *Bowley*, Elements of Statistics, 5. vyd. Londýn 1920, *Edgeworth*, heslo Index-numbers v *Palgrave's Dictionary of Political Economy* II., *Olivier*, Les nombres-indices de la variation des prix, Paříž 1927.

## ŘADY ČASOVÉ.

### 1. Charakteristika a vyrovnávání časových řad.

Soubory statistické, které jsou upraveny v časovém sledu statistických údajů podle obvyklých jednotek časových, t. zv. řady časové, vykazují zvláštní vlastnosti, kterými se odlišují od jiných souborů statistických, a které vyžadují zvláštních method analytických.

Časové řady představují podle své povahy časový vývoj nějakého souboru. Tento vývoj může míti různý průběh, který lze zařaditi do několika základních typů.

Ponecháváme-li stranou takové statistické řady časové, jejichž pohyb od hodnoty k hodnotě je zcela nepravidelný, můžeme rozezná vati tyto hlavní typy:

1. Typ řad stálých (konstantních), které probíhají toliko s malým kolísáním hodnot, jako je na př. časová řada rozdělení pohlaví u novorozенých dětí (v. př. č. 2).

2. Typ řad vývojových (evolučních), ve kterých hodnoty vykazují stálý vzestup nebo stálý sestup, na př. řada úmrtnosti, natality, vykazovaly v letech předcházejících druhé světové válce trvalý pokles, řada počtu obyvatelstva, řada počtu automobilů a j., v téže době trvalý vzestup.

3. Typ řad periodických, které vykazují vzestup a sestup hodnot v určitých obdobích se opakující. Nejlépe vynikne tento pohyb při grafickém znázornění řady, kde čára znázorňující řadu vykazuje periodické klesání a stoupání v podobě vln, na př. řada vyjadřující těžbu uhlí, přistavování nákladních vagonů, a jiné řady souborů hospodářské povahy.

Vývojové řady mohou nabývati různých podob, z nichž některé jsou opět charakteristické. zvláště, znázorníme-li je graficky. Jedním z těchto



případů je vzrůst řady v podobě paraboly, kde přírůstky řady jsou nejprve velké, pak se zmenšují, takže časová řada se pak poněkud obrací ve směr vodorovný.

Jestliže naopak přírůstky stále stoupají, má časová řada tvar prudce stoupající křivky, podobající se písmenu *J*. Jestliže pak přírůstky nejprve stoupají a pak od určité doby klesají, dostáváme řadu, která nejprve stoupá pomalu, pak rychle a na konci opět nabývá povahy stálé. Tato křivka se označuje též jako křivka logistická, a vyskytuje se často v soubořích jevů demografických a hospodářských, na př. vzrůst obyvatelstva, peněžních vkladů u spořitelny a j., děje se v této podobě (stejnou podobu měla t. zv. křivka kumulační, znázorňující rozložení četností při stálém přičítání četností, srv. př. č. 11).

Podobně jako u řad vzestupných lze i u řad sestupných nalézt podobné typy, kde se vyskytují obdobné pravidelnosti v úbytcích hodnot.

Byla nalezena řada method, které se snaží vyrovnati časové řady empirické na podoby (tvary), které odpovídají některé theoretické formě (na př. parabole), jako byly řady nahoře uvedené.

Methody, jichž se nejčastěji užívá k vyrovnání časových řad, jsou:

1. *metoda pohybujících průměrů (klouzavých průměrů)*. Časté použití její je ve statice konjunktury hospodářské, máme-li řadu hodnot podléhajících vlivům sezónním (t. j. změnám průběhem roku, vzniklým vlivem střídání ročních období), a chceme-li řadu od těchto vlivů sezónních oprostiti. U takové časové řady, vyjadřující měsíční hodnoty, na př. měsíční součty vývozu a dovozu, počet přistavených vagonů, počet nezaměstnaných k určitému dni každého měsíce a pod., vypočítávají se při této methodě postupně průměry za dvanáct měsíců, při čemž se přihlíží pouze k polovině hodnoty prvního měsíce a připojí se polovina hodnoty měsíce třináctého (t. j. prvního měsíce následujícího roku) a tak postupně podle schématu:

Jsou-li  $m_1, m_2, m_3, m_4, m_5, m_6, \dots, m_{12}$  měsíční hodnoty jednoho roku,  $m_{13}, m_{14}, m_{15}, m_{16}, \dots, m_{24}$  měsíční hodnoty druhého roku, vznikne pak řada, jejíž první člen je průměr:

$$\frac{\frac{1}{2} m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + m_9 + m_{10} + m_{11} + m_{12} + \frac{1}{2} m_{13}}{12}$$

pro střední měsíc, t. j. červenec, druhý člen je průměr:

$$\frac{\frac{1}{2} m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + m_9 + m_{10} + m_{11} + m_{12} + m_{13} + \frac{1}{2} m_{14}}{12}$$

pro měsíc srpen, a tak postupně dále.

Řada z těchto průměrů vytvořená zobrazuje pak vývoj konjunktury, zbařený sezónních kolísání.

Příklad č. 19. Mezinárodní indexy nezaměstnanosti 1929—1932.

Měsíc	(1929 = 100.)							
	Řada bez vyloučení sezónních výkyvů				Řada s vyloučením sezónních výkyvů			
	1929	1930	1931	1932	1929	1930	1931	1932
Leden . . . . .	128	147	233	287	89	118	195	256
Únor . . . . .	131	153	235	289	89	125	200	261
Březen . . . . .	103	146	228	281	89	131	206	266
Duben . . . . .	86	141	214	276	90	138	211	269
Květen . . . . .	75	138	203	272	92	146	217	273
Červen . . . . .	70	138	202	269	95	154	224	276
Červenec . . . . .	72	148	208	277	97	162	230	279
Srpen . . . . .	73	155	215	274	99	169	234	280
Září . . . . .	74	157	221	269	100	175	237	280
Říjen . . . . .	85	164	227	266	103	180	241	280
Listopad . . . . .	98	183	243	275	107	185	245	279
Prosinec . . . . .	124	209	268	289	112	190	251	278
Indexy . . . . .	100	168	241	297				

2. metoda vyrovnávání řad cestou grafickou (srv. kap. 13.);

3. metoda nejmenších čtverců. Metoda tato záleží v tom, že se zvolí čára, která se přizpůsobí dané statistické řadě (empirické křivce) do té míry, že součet dvojmočí (čtverců) odchylek od této křivky je minimum, t. j. veličina menší než součet dvojmočí odchylek od jakékoli křivky jiné.

Nejjednodušším případem takové čáry je přímka; může to být však též parabola, exponenciální křivka a j. Podmínky stanovené k určení potřebných konstant pro rovnici této čáry jsou: aby součet odchylek empirických hodnot (které si představujeme v grafickém znázornění jako body sestavené na dvou kolmých osách, srv. kap. 13.) od příslušných bodů hledané čáry byl roven nule, a aby součet čtverců těchto odchylek byl minimem.

Zvolili jsme jako nejjednodušší případ přímku, jejíž rovnice zní  $y = a + bx$ ; je tedy třeba zjistiti konstanty  $a$  a  $b$  v této rovnici. Rovnice potřebné pro výpočet těchto konstant, v tomto případě dvě, zvané rovnicemi normálními, získají se sečtením všech rovnic přímky pro všechny zvolené hodnoty podle vzorce:

$$\text{I. normální rovnice } \Sigma(y) = na + b\Sigma(x)$$

$$\text{II. normální rovnice } \Sigma(xy) = ax + b\Sigma(x^2).$$

při čemž  $n$  znamená opět počet hodnot řady, a tedy též počet sestavených rovnic.

Příklad č. 20. Časová řada vyjadřující počet obyvatelstva v Německu v letech 1871—1910.<sup>1)</sup>

1. prosince 1871	41,06 mil. obyvatel
1. prosince 1880	45,23 „ „
1. prosince 1890	49,43 „ „
1. prosince 1900	56,37 „ „
1. prosince 1910	64,93 „ „

Střední hodnotu této časové řady volíme tak, aby vyhovovala podmínce, že součet odchylek se rovná nule (jako tomu bylo u aritmetického průměru), tedy hodnotu pro r. 1890.

$$41,06 = a - 19b$$

$$45,23 = a - 10b$$

$$49,43 = a$$

$$56,37 = a + 10b$$

$$64,93 = a + 20b$$

---


$$257,02 = 5a + b = \text{I. normální rovnice.}$$

Druhou normální rovnici dostaneme, když násobíme každou z foregoing rovnic příslušnou hodnotou  $x$ :

$$- 780,14 = - 19a + 361b \quad (x = - 19)$$

$$- 452,30 = - 10a + 100b \quad (x = - 10)$$

$$563,70 = 10a + 100b \quad (x = 10)$$

$$1298,60 = 20a + 400b \quad (x = 20)$$

---


$$629,86 = a + 961b = \text{II. normální rovnice.}$$

Z obou normálních rovnic vyplývá

$$a = 51,284$$

$$b = 0,60206$$

Rovnice vyrovnané přímkou pak zní

$$y = 51,284 + 0,60206x.$$

Hodnoty podle této rovnice vyrovnané pro jednotlivé roky časové řady jsou (dosadíme příslušné hodnoty veličiny  $x$ ):

Rok 1871	39,85 mil.
1880	45,26 „
1890	51,28 „
1900	57,31 „
1910	63,34 „

To jsou theoretické hodnoty za předpokladu, že časová řada počtu obyvatelstva odpovídá rovnici přímkou.

<sup>1)</sup> Winkler, Grundriß der Statistik, Berlin 1931, I. str. 105.



## 2. Interpolace časových řad.

Methody pro vyrovnávání řad mohou sloužiti též k výpočtu chybějících hodnot statistické řady (t. zv. *interpolace*) nebo k odhadu neznámých, budoucích hodnot řady (t. zv. *extrapolace řady*).

Ze známých hodnot statistické řady doplňujeme tedy řadu vypočítanými hodnotami za ty, které neznáme. Jestliže v řadě časové chybí jen některý člen (některá hodnota) řady, lze při interpolaci hodnot postupovati též jednoduššími methodami než je methoda nejmenších čtverců.

Na př. sčítání lidu se koná u nás pouze jednou za deset let, máme však zájem na tom znáti hodnoty časové řady vyjadřující počet obyvatelstva též pro roky, kdy se sčítání lidu nekonalo (potřebujeme znáti počet obyvatelstva pro výpočet poměrných čísel úmrtnosti, natality a j., jimž za základ slouží počet obyvatelstva v každém roce). Interpolace této hodnoty se může státi dvojím způsobem, jak patrnó z následujícího příkladu.

**Příklad č. 21.** Známe počet obyvatelstva v Československu r. 1921 (sčítání se konalo 15. února 1921) 13,612.424 obyvatel, a rovněž r. 1930 (sčítání konalo se 1. prosince 1930) 14,726.158 obyvatel. Předpokládáme-li, že vzrůst obyvatelstva mezi těmito roky se děl pravidelně každý rok o stejné číslo, čili že obyvatelstva přibývalo řadou aritmetickou, dělíme rozdíl mezi oběma hodnotami devíti a dostaneme tak roční přírůstek obyvatelstva (rovná se = 123.748). Interpolovaná hodnota počtu obyvatelstva na př. pro r. 1926 rovná se pak hodnotě základní, zvětšené o pateronásobný přírůstek roční

$$13,612.424 + 5 \times 123.748 = 14,231.164.$$

Předpokládáme-li však, že *poměrný* (nikoli absolutní, jako dříve) přírůstek obyvatelstva byl každý rok stejný, že tedy obyvatelstvo vzrůstalo řadou geometrickou, musili bychom při interpolaci hodnoty obyvatelstva pro r. 1926 postupovati podle vzorce analogického pro výpočet vzrůstu kapitálu při složitém úrokování (připočítávání úroků z úroků):

$b = aq^n$ , kde  $b$  značí počet obyvatelstva r. 1930,  $a$  počet obyvatelstva r. 1921,  $n$  počet roků. Z této rovnice vyplývá, že

$$q = \sqrt[n]{\frac{b}{a}}.$$

Dosadíme-li uvedené hodnoty, dostaneme (pro umocňování a odmocňování je nejlépe použití logaritmů):

$$q = \sqrt[9]{\frac{14,726.158}{13,612.424}} = 1,00878.$$

$$\begin{aligned}\text{Obyvatelstvo r. 1926 se pak rovná} &= 13,612,424 \times q^5 = \\ &= 13,612,424 \times 1,00878^5 = 14,220,350.\end{aligned}$$

Počet obyvatelstva interpolovaný pro r. 1926 za předpokladu, že obyvatelstvo vzrůstá každoročně o poměrný stejný přírůstek čili při vzrůstu obyvatelstva čím dále tím o větší čísla absolutní, jest ovšem o něco menší než interpolovaná hodnota za předpokladu, že vzrůstá každoročně o stejné číslo absolutní.<sup>1)</sup>

K interpolaci může se použítí též metody grafické, křivka znázorňující statistickou řadu a pevně určená známými hodnotami řady umožňuje nám zjistiti též členy neznámé.

Při používání hodnot, získaných interpolací, nesmí se zapomenouti, že nejsou to hodnoty skutečně získané statistickým šetřením, nýbrž pouze hodnoty vypočítané za určitých předpokladů a tedy méně spolehlivé; jest zřejmé, že na př. v uvedeném příkladě mohla býti pravidelnost vzrůstu obyvatelstva během desítiletí porušena, a potom interpolovaná hodnota neodpovídá skutečnosti.

Úsudek na budoucí hodnoty časových řad statistických, t. j. výpočet očekávaných hodnot časové řady v budoucích obdobích časových se nazývá *extrapolací*. Extrapolovati hodnoty je však možno s jistou pravděpodobností pouze pro krátká období časová, neboť předpoklad, že vývoj v budoucnosti se bude pohybovati stejným způsobem jako v pozorované době minulé, stává se tím nejistějším, čím vzdálenější je toto budoucí období.

Zvláštní význam mají řady časové ve statistice hospodářské, na jejich rozboru se zakládá zejména statistika hospodářské konjunktury.

**Literatura ke kap. 9.** Vedle spisů v textu citovaných *Mráz: Statistické řady*, *Sborník věd právních a státních XXXII*, 1932.

<sup>1)</sup> Při interpolaci hodnot nebylo přihlíženo k okolnosti, že sčítání r. 1921 stalo se 15. února, sčítání druhé 1. prosince. Mezi oběma sčítáními není tedy doba 9letá, nýbrž 9 roků a 9 ½ měsíce, čili 9,7917 let. Podle toho byl by správnější přírůstek při lineární interpolaci 113.743 a počet obyvatel pro 15. únor 1926 14,181.139. Při geometrické interpolaci by se pak  $n$  rovněž rovnalo  $n = 9,7917$ . Pak by se  $q$  rovnalo  $q = 1,00806$ . Interpolovaný počet obyvatelstva pro 15. únor 1926 byl by pak 14,170.209.

## STATISTICKÉ ZÁVISLOSTI.

### 1. Nomologický úkol statistiky.

Vedle popisného úkolu má statistická metoda úkol analytický nebo přesněji nomologický či nomothetický (od řeckého slova *nomos*, zákon) který se neomezuje na popisování a charakterisování souborů a na srovnávání statistických údajů, nýbrž hledá též vnitřní závislosti mezi statistickými soubory a jejich znaky. Srovnáváme-li počet obyvatelstva dvou států, vyjádřený buď v číslech absolutních nebo relativních (hustotu obyvatelstva), statistické charakteristiky dvou souborů atd., jde v podstatě stále o úkol popisný, naproti tomu zjišťujeme-li vnitřní vztahy mezi národností a náboženským vyznáním v souboru obyvatelstva, mezi pracovní dobou a mzdou v souboru dělnictva, jde o vlastní úkol nomologický. Tento druhý úkol statistiky je ještě významnější pro toho, kdo se zabývá studiem věd sociálních než úkol popisný.

Vzhledem k zvláštní logické povaze statistických souborů a statistické metody nejde ve statistice o hledání závislostí funkčních (poměr funkční znamená v matematice, že každé hodnotě veličiny jedné odpovídá určitá hodnota veličiny druhé — nezávisle proměnné závisle proměnná), nýbrž o závislost zvláštního druhu, závislost statistickou nebo kolektivní. Někdy se mluví též o stochastické závislosti mezi jevy, která se zjišťuje statistickou metodou (od řeckého slova *stochazesthai*, domnívati se, tušiti). Statistická nebo stochastická závislost je ovšem volnější než závislost funkční nebo závislost kausální, neboť se zakládá na pouhém poměru pravděpodobnosti, nikoli poměru nutnosti. Těže povahy jsou pak i jednotlivé míry (koeficienty) statistických závislostí.

Při hledání závislostí mezi kolektivními jevy a jejich znaky je třeba



používati obecných pravidel logických vedle zvláštních method, které poskytuje statistika. Dále je třeba vždy míti na paměti, že ten, kdo chce pracovati statistickou methodou v některém oboru sociálních jevů, musí znáti nejen tuto methodu, nýbrž též příslušný obor věd, na př. pracuje-li statistickou methodou v oboru jevů hospodářských, znáti též základy národního hospodářství.

## 2. Závislost mezi znaky kvalitativními.

Zjistíme-li u nějakého souboru při statistickém šetření dva znaky kvalitativní, které označíme písmeny  $A$  a  $B$  (na př. národnost česká a náboženské vyznání římsko-katolické), můžeme zkoumati, jsou-li tyto znaky nezávislé či závislé. Jsou-li vnitřně nezávislé, znamená to, že v souboru existence (přítomnost) znaku  $A$  nemá vlivu na existenci (přítomnost) znaku  $B$ , přesněji řečeno, že pravděpodobnost znaku  $A$  nemění se existencí znaku  $B$  a naopak. Není-li tomu tak, nýbrž pravděpodobnost znaku  $A$  se zvětšuje nebo zmenšuje přítomností znaku  $B$  (poměr může býti také opačný), mluvíme o kolektivní nebo stochastické závislosti mezi znaky kvalitativními. Při tom může to býti buď přímá závislost, jestliže pravděpodobnost znaku jednoho se zvětšuje přítomností znaku druhého, nebo nepřímá závislost (záporná asociace), jestliže pravděpodobnost znaku jednoho se zmenšuje přítomností znaku druhého.

Pro zkoumání závislosti znaků kvalitativních jest několik kritérií. Všechna se v podstatě zakládají na srovnávání *rozdílů* mezi poměrnými četnostmi dílčích souborů utvořených podle různých znaků kvalitativních nebo podle různých obměn téhož znaku kvalitativního (methoda *rozdílů* nebo *diferenční*). Při zkoumání závislosti znaků kvalitativních používáme symbolů navržených anglickým statistikem *Yulem*.<sup>1)</sup>

Písmeny  $A, B$  označíme přítomnost znaků (znaky pozitivní);

písmeny  $a, b$  označíme nepřítomnost znaků (znaky negativní);

písmeny  $(A), (B)$  v závorce označíme počet (četnost) případů, vykazujících znaky pozitivní;

písmeny  $(a), (b)$  v závorce označíme počet (četnost) případů, vykazujících znaky negativní;

písmeny  $A_1, A_2, A_3$  atd. označíme obměny znaku  $A$ , písmeny  $(A_1), (A_2), (A_3)$  v závorce označíme opět analogicky četnost případů, vykazujících příslušné obměny znaku  $A$ .

Písmenem  $N$  označuje se pak počet všech případů (celý soubor).

<sup>1)</sup> *Yule, An introduction to the Theory of Statistics* 11 vyd. Londýn 1937, čes. překlad 7. vyd. Úvod do theorie statistiky, Praha 1926, kap. I—V.

Kriterium nezávislosti mezi znaky  $A$  a  $B$  vyjadřuje se pak rovnicemi:

$$\frac{(A B)}{(B)} = \frac{(A \beta)}{(\beta)},$$

t. j. poměr počtu případů, kde vyskytuje se znak  $A$ , tak  $B$  pozitivně, ke všem případům, v nichž se vyskytuje  $B$ , musí být roven poměru počtu případů, kde vyskytuje se znak  $A$  pozitivní, a znak  $B$  není přítomen, k počtu všech případů, kde znak  $B$  není přítomen; nebo rovnicí v jiné formě:

$$(A B) = \frac{(A) (B)}{N},$$

t. j. počet případů, v nichž vyskytuje se jak znak  $A$ , tak znak  $B$ , musí být roven poměru součinu případů, kde vyskytuje se znak  $A$ , a počtu případů, kde se vyskytuje znak  $B$ , k počtu všech případů; nebo rovnicí t. zv. zkřížených znaků:

$$(A B) (\alpha \beta) = (\alpha B) (A \beta),$$

t. j. součin případů, v nichž vyskytuje se znak  $A$  i  $B$  a případů, v nichž nevyskytují se ani znak  $A$  ani  $B$ , musí se rovnat součinu případů, v nichž vyskytuje se znak  $B$ , ale nikoli  $A$ , a případů, v nichž vyskytuje se znak  $A$ , ale nikoli  $B$ .

Z těchto rovnic dají se odvoditi postupným dosazováním ještě některé rovnice jiné, kterými lze rovněž zkoumati nezávislost znaků  $A$  a  $B$ .<sup>1)</sup>

<sup>1)</sup> Z první nahoře uvedené rovnice dají se postupným dosazováním odvoditi rovnice:

$$\begin{aligned} \frac{(a B)}{(B)} &= \frac{(a \beta)}{(\beta)} \\ \frac{(A B)}{(A)} &= \frac{(u B)}{(u)} \\ \frac{(A \beta)}{(A)} &= \frac{(a \beta)}{(a)} \end{aligned}$$

z druhé nahoře uvedené rovnice dají se postupným dosazováním odvoditi rovnice:

$$\begin{aligned} \frac{(A B)}{(B)} &= \frac{(A)}{N} \\ \frac{(A B)}{(A)} &= \frac{(B)}{N} \\ \frac{(A B)}{N} &= \frac{(A)}{N} \times \frac{(B)}{N} \end{aligned}$$

Tuto poslední rovnici definuje Yule pravidlem: Jsou-li znaky  $A$  a  $B$  na sobě nezávislé, pak poměrná četnost současného se vyskytnutí znaků  $A B$  je rovna součinu poměrných četností jednotlivých znaků  $A$  a  $B$ .

Toto pravidlo je založeno na větě o násobení pravděpodobností, neboť matematická pravděpodobnost, že  $A$  a  $B$  se zároveň vyskytnou, rovná se součinu jednotlivých pravděpodobností, jsou-li jevy na sobě nezávislé (viz str. 21).

Nejsou-li však znaky  $A$  a  $B$  nezávislé, nýbrž je-li mezi nimi závislost, vznikají na místě rovnice nerovnice, které ukazují, jaká závislost mezi oběma znaky je dána:

Jestliže  $(A|B) > \frac{(A)(B)}{N}$ , pravíme, že  $A$  a  $B$  jsou závislé přímo (asociovány kladně),

jestliže  $(A|B) < \frac{(A)(B)}{N}$ , pravíme, že  $A$  a  $B$  jsou závislé nepřímo (záporně sdruženy, disasociovány).

Přímá závislost mezi dvěma znaky  $A$  a  $B$  znamená, že počet jevů vykazujících znak  $A$  a zároveň znak  $B$ , převyšuje počet, jaký bychom očekávali, kdyby znaky  $A$  a  $B$  byly nezávislé; nepřímá závislost pak, že počet jevů, vykazujících znak  $A$  a zároveň znak  $B$ , jest menší, než jaký bychom očekávali, kdyby znaky  $A$  a  $B$  byly nezávislé.

Při úsudcích na přímou nebo nepřímou závislost mezi dvěma znaky jest však třeba zachovati určitou opatrnost. Jestliže stupeň (intensita) asociace jest malý, na př. při uvedeném vzorci levá strana liší se od pravé jen o malé hodnoty (v poměru k počtu pozorovaných případů), nelze ještě bezpečně usuzovati na závislost obou znaků. Tuto věc lze vysvětliti theoreticky tím, že na asociaci působí složitá řada mnoha příčin, jejichž podrobné působení neznáme, a o jejichž výsledku pravíme, že byl způsoben nahodilými okolnostmi nebo kolísáním náhodných výběrů<sup>1)</sup> (jinými slovy náhodou při výběru souboru).

Jako příklad takové zdánlivé závislosti uvádí Yule<sup>2)</sup> soubor, skládající se ze zaznamenaných vrhů mincí, při čemž znakem kvalitativním je výsledek hodu, zda padl rub či líc.

#### Příklad č. 22.

Při 100 pozorovaných dvojic vrhů bylo:

Při prvním vrhu líc a druhém rub . . . . . 18

Při prvním vrhu líc a druhém rovněž líc . . . . . 26

Při prvním vrhu rub a při druhém líc . . . . . 27

Při prvním vrhu rub a při druhém rovněž rub . . . . . 29

Jestliže označíme písmenem  $A$  líc při prvním vrhu a písmenem  $B$  líc při druhém vrhu, máme četnost znaků

$$(A) = 44, (B) = 53.$$

<sup>1)</sup> Yule označuje to jako „fluctuations of sampling“, l. c. str. 31.

<sup>2)</sup> Yule l. c. str. 30.



$$\frac{(A)(B)}{N} = \frac{44 \times 53}{100} = 23,32, (A B) = 26.$$

$$(A B) \text{ je tedy větší } > \frac{(A)(B)}{N},$$

z čehož bychom mohli usuzovati na přímou závislost mezi oběma znaky, t. j. mezi výsledkem prvního hodu a výsledkem druhého hodu, ačkoli pro takovou závislost není žádného předpokladu. Takovéto výsledky, zdánlivé závislosti, považují se za důsledek nahodilých okolností nebo kolísání nahodilých výběrů (pozorování 100 dvojic vrhů možno považovati za nahodilý výběr z velkého souboru vrhů, ve kterém mohou se uplatňovati vlivy náhodné, na př. v uvedeném příkladě způsob, jakým vrháme minci a pod.).

Intensitu závislosti můžeme vedle vzorce uvedeného zkoumati na základě rozdílu poměrných čísel, vyjádřených v procentech nebo promile podle vzorců:

$$\begin{array}{ll} \frac{(A B)}{(B)} > \frac{(A)}{N} & \frac{(A B)}{(A)} > \frac{(B)}{N} \\ \frac{(A B)}{(B)} > \frac{(A \beta)}{(\beta)} & \frac{(A B)}{(A)} > \frac{(\alpha B)}{(\alpha)} \end{array}$$

atd.

Ze vzorců nejvýhodnější pro praktické použití jsou ty, které jednak ukazují jasněji stupeň závislosti, jednak osvětlují důležitější stránku zkoumaného problému podle povahy statistického materiálu.

**Příklad č. 23. Závislost mezi hluchoněmostí a slabomyslností na základě sčítání lidu v Anglii a Walesu r. 1901.<sup>1)</sup>**

Úhrnný počet obyv. Anglie a Walesu . . . . .	32,528.000 (N)
Počet slabomyslných . . . . .	48.882 (A)
Počet hluchoněmých . . . . .	15.246 (B)
Počet slabomyslných a hluchoněmých . . . . .	451 (AB)

Znak slabomyslnosti označme písmenem A, znak hluchoněmosti písmenem B. Vycházíme-li od znaku hluchoněmosti B, t. j. chceme-li především zjistiti poměry mezi hluchoněmými, použijeme vzorce  $\frac{(A B)}{(B)} > \frac{(A)}{N}$

$$\frac{451}{15.246} > \frac{48.882}{32.528.000}, \text{ vyjádřeno v promile } 29,6\text{‰} > 1,5\text{‰}.$$

<sup>1)</sup> Yule l. c. str. 33.

Vycházíme-li od znaku slabomyslnosti  $A$ , t. j. chceme-li především zjistiti poměry mezi slabomyslnými, použijeme vzorce

$$\frac{(A \ B)}{(A)} > \frac{(B)}{N}$$

$$\frac{451}{48.882} > \frac{15.246}{32,528.000}, \text{ vyjádřeno v promile } 9,2^0/_{00} > 0,5^0/_{00}.$$

Srovnání závislosti znaků jest v obojím případě velmi názorné a ukazuje značný stupeň závislosti.

**Příklad č. 24. Závislost mezi pohlavím a úmrtností.**

Podle sčítání lidu z r. 1930 bylo v Čechách r. 1930

mužů	3,452.123
žen	3,657.253
úhrnem všeho obyvatelstva	7,109.376
zemřelo téhož roku	
mužů	49.398
žen	46.199
úhrnem zemřelo	95.597.

Označíme-li pohlaví mužské písmenem  $A$ , úmrtí písmenem  $B$ , podává se pro srovnání závislosti mezi pohlavím a úmrtími jako nejvýhodnější vzorec, který přímo ukazuje poměrnou četnost zemřelých mužů v souboru mužů vůbec a zemřelých žen v souboru všech žen.

$$\frac{(A \ B)}{(A)} > \frac{(\alpha \ B)}{(\alpha)}$$

$$\frac{49.398}{3,452.123} > \frac{46.199}{3,657.253}$$

vyjádřeno v promile (na tisíc obyvatel)

$$14,30^0/_{00} > 12,62^0/_{00}.$$

je tedy dána přímá závislost mezi pohlavím mužským a úmrtími.

Srovnáme-li úmrtnost mezi muži s úmrtností mezi veškerým obyvatelstvem podle vzorce

$$\frac{(A \ B)}{(A)} > \frac{(B)}{N},$$

kde  $(B)$  značí veškerá úmrtí v Čechách r. 1930 (95.597) a  $N$  veškeré obyvatelstvo Čech (7,109.376), dostaneme nerovninu

$$14,30^0/_{00} > 13,44^0/_{00}.$$

Také tento poměr ukazuje, že oba znaky jsou závislé, je však méně výrazný než nahoře uvedený.

Pochybnost, zda rozdíl mezi poměrnými četnostmi ukazuje skutečnou závislost mezi znaky, či zda je pouze náhodný, je možno řešiti pomocí výpočtu theoretické střední odchylky rozdílu mezi poměrnými četnostmi. Jestliže rozdíl mezi poměrnými četnostmi převyšuje aspoň trojnásobek theoretické střední odchylky, považuje se obvykle nahodilost za vyloučenou a rozdíl takový za spolehlivého ukazatele závislosti mezi znaky.

Pro výpočet theoretické střední odchylky rozdílu platí vzorec:

$$\sigma_d = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}},$$

t. j. rovná se odmocnině součtu čtverců theoretických středních odchylek obou poměrných četností. Matematickou pravděpodobnost  $p$  a  $q$  nahradíme opět pravděpodobností empirickou, takže dostaneme vzorec:

$$\sigma_d = \sqrt{\frac{\frac{m_1}{n_1} \left(1 - \frac{m_1}{n_1}\right)}{n_1} + \frac{\frac{m_2}{n_2} \left(1 - \frac{m_2}{n_2}\right)}{n_2}}.$$

V uvedeném příkladě č. 24 dosadíme hodnoty:

$$\sigma_d = \sqrt{\frac{0,01430 \times 0,98570}{3,452.123} + \frac{0,01262 \times 0,98738}{3,657.253}} = 0,0000864.$$

Trojnásobek této theoretické střední odchylky  $3 \sigma_d = 0,0002592$ . Rozdíl poměrných četností je větší než trojnásobek střední odchylky  $1,68^0_{00} = 0,00168 > 0,0002592$ .

Rozdíl mezi poměrnými četnostmi, ač není velký, můžeme tedy považovati za spolehlivé kritérium přímé závislosti mezi pohlavím mužským a úmrtností.

K měření velikosti závislosti (asociace) mezi znaky používá Yule zvláštního vzorce, který nazývá koeficientem asociace a označuje písmenem  $Q$  podle začátečního písmena jména *Quételetova*. Poněvadž písmenem  $Q$  se označuje rovněž divergenční koeficient *Lexisův*, o němž bude řeč v kapitole o náhodných výběrech, používáme na rozlišení od tohoto koeficientu písmena  $Q$  s indexem *as* (asociace).

Tento koeficient asociace se rovná:

$$Q_{as} = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

(v čitateli je rozdíl, ve jmenovateli součet zkřížených znaků).

Jeden krajní případ nastává, jestliže  $(A\beta)(\alpha B) = 0$ , potom se koeficient rovná  $Q_{as} = 1$ . Tento případ se nazývá úplnou asociací, a jsou to případy, kdy všechna  $A$  jsou  $B$  nebo naopak.



Druhá mez je případ, jestliže  $(AB) (\alpha\beta) = 0$ , potom koeficient rovná se  $Q_{as} = -1$ . Tento případ se nazývá úplnou disasociací, a jsou to případy, kdy žádné  $A$  není  $B$  nebo naopak.

Koeficient asociace může býti považován za míru závislosti dvou znaků kvalitativních; čím více blíží se  $+1$ , tím je přímá závislost vyšší, čím více blíží se  $-1$ , tím je nepřímá závislost vyšší; blíží-li se nule, nelze usuzovati na závislost. To vše platí opět s výhradou nahodilých shod, které se mohou projevit zejména při menším počtu pozorování.<sup>1)</sup>

V příkladě výše uvedeném o závislosti mezi hluchoněmostí a slabomyslností na podkladě statistických údajů ze sčítání lidu v Anglii a Walesu (v. př. č. 23) rovná se koeficient asociace  $Q_{as} = +0,91$ , je tedy značně vysoký.

**Příklad č. 25. Závislost mezi mužským pohlavím a bezkonfesností.**

Při sčítání lidu r. 1921 zjištěno bylo v Československu mužů bez vyznání 417.319, žen bez vyznání 307.188, všech mužů bylo 6,559.503, všech žen bylo 7,053.669; tážeme se, jakého stupně je závislost mezi mužským pohlavím a bezkonfesností?

Označíme písmenem  $A$  pohlaví mužské, písmenem  $B$  bezkonfesnost; potom dostaneme četnosti:

$$(A) = 6,559.503, \quad (\alpha) = 7,053.669,$$

$$(AB) = 417.319, \quad (\alpha B) = 307.188.$$

Četnosti ve vzorci ještě potřebné snadno odvodíme dosazením ze známých četností:

$$(A\beta) = (A) - (AB) = 6,559.503 - 417.319 = 6,142.184,$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = 7,053.669 - 307.188 = 6,746.481.$$

$$\begin{aligned} Q_{as} &= \frac{(A B) (\alpha \beta) - (A \beta) (\alpha B)}{(A B) (\alpha \beta) + (A \beta) (\alpha B)} = \\ &= \frac{(417.319 \times 6,746.481) - (6,142.184 \times 307.188)}{(417.319 \times 6,746.481) + (6,142.184 \times 307.188)} = \\ &= \frac{928.629,485.847}{4,702.239,923.031} = 0,197. \end{aligned}$$

<sup>1)</sup> Pro výpočet stupně závislosti mezi znaky kvalitativními používá se ještě jiných vzorců, jako je *Pearsonův* koeficient kontingence a *Giniho* koeficient podobnosti, avšak žádný z uvedených vzorců nelze považovati za tak přesný, aby číslo koeficientu mohlo býti považováno za přesvědčivý stupeň asociace. O tom, jak naopak tyto vzorce mohou vésti k nesprávným výsledkům a jak jsou závislé na zastoupení znaků v souboru srv. *Boháč, Zdáni a skutečnost ve statistice*, Praha 1935. Pro statistickou praxi možno se spokojiti hrubší methodou diferenční, nahoře vyličenou.

Je tedy z příkladu patrná přímá závislost mezi znakem „mužské pohlaví“ a znakem „bezkonfesnost“ v souboru obyvatelstva, stupeň závislosti není však velký.

Dělí-li se znak kvalitativní na více než dvě obměny, jako je tomu na př. při znaku národnosti, náboženského vyznání a pod., mluvíme o množném třídění znaků na rozdíl od třídění dvojdílného či dichotomického, jako byly případy probírané.

V takových složitých případech postačí pro běžnou statistickou praxi vyjádření poměrné četnosti pro jednotlivé obměny znaku jednoho v kombinaci s příslušnými obměnami znaku druhého a srovnávatí opět rozdílů mezi poměrnými četnostmi.

Takové srovnání děje se obyčejně úpravou souboru do tabulky podle obměn obou znaků.

**Příklad č. 26. Znalost čtení a psaní podle národnostních skupin.**

**Čechy r. 1930:**

Národnostní skupina	z 1000 mužů starších deseti let neumělo ani číst ani psát	z 1000 žen starších deseti let neumělo ani číst ani psát
československá	9,8	16,6
německá	7,9	11,1
ze všeho obyvatelstva	9,7	14,9

**Slovensko r. 1930:**

československá	62,6	91,5
ruská (ukrajinská)	245,3	341,7
německá	35,4	50,3
maďarská	44,1	55,6
ze všeho obyvatelstva	67,5	94,5

Poměrné četnosti můžeme srovnávatí jednak mezi jednotlivými skupinami, jednak s poměrnými četnostmi pro veškeré obyvatelstvo. Kdežto v Čechách rozdíly mezi poměrnými četnostmi mezi skupinou československou a německou nebyly příliš výrazné, byly na Slovensku tyto rozdíly zřejmé a ukazují závislost mezi příslušností k ruské (ukrajinské) národnosti a částečně též československé národnosti a negramotností. Rovněž se objevuje závislost mezi ženským pohlavím a negramotností a to u všech národnostních skupin a ovšem i v úhrnu obyvatelstva. Ovšem v úsudech je třeba opatrnosti, neboť závislosti ty mohou býti do jisté

míry zprostředkovány znaky jinými, na př. velká většina příslušníků ruské (ukrajinské) národnosti na Slovensku žije ve venkovských obcích a živí se zemědělstvím a lesnictvím, závislost může tedy být ve velké míře zprostředkována znaky místního usídlení a příslušnosti k povolání.

Zdánlivá asociace bývá často způsobována t. zv. asociací dílčí (parciální), t. j. tím, že závislost mezi dvěma znaky zprostředkována je závislostí na znaku třetím. Na př. při celkovém srovnávání úmrtnosti příslušníků různých povolání může být závislost mezi úmrtností a některým povoláním způsobována vlivem znaku třetího, t. j. věku příslušníků těchto povolání, který je různý, poněvadž vyšší věkové třídy jsou zastoupeny u jednoho povolání četněji než u druhého a důsledkem toho je též jejich vyšší úmrtnost. Závislost mezi znaky je tedy způsobována dílčí asociací se znakem třetím, povahy kvantitativní, věkem.

Je-li takový další znak obsažen ve statistických údajích, je třeba při zjišťování závislosti mezi znaky též k němu přihlížeti tím, že roztrídíme statistické údaje nejprve podle tohoto třetího znaku  $C$ , jinak aspoň pamatovati na to, zda zjištěná závislost nemůže takovým způsobem vzniknouti.

Vzorec platný pro zjištění závislosti dvou znaků  $A$  a  $B$  v souboru případů  $C$  zní:

$$(ABC) > \frac{(AC)(BC)}{(C)}, \text{ je-li závislost přímá, a}$$

$$(ABC) < \frac{(AC)(BC)}{(C)}, \text{ je-li závislost nepřímá.}$$

Obdobným způsobem, jako jsme usuzovali na závislost mezi obměnami dvou znaků kvalitativních z rozdílů poměrných četností, můžeme postupovati též při posuzování závislosti mezi dvěma znaky kvalitativními (v následujícím příkladu č. 27 mezi úmrtností a legitimitou), při čemž však celý soubor je rozdělen ve skupiny podle znaku kvantitativního (věku), čímž je umožněno jemnější srovnávání v jednotlivých skupinách.

Uvádíme příklad dánského statistika *Westergaarda*<sup>1)</sup> o zkoumání závislosti mezi úmrtností dětí a jejich legitimitou, poněvadž dobře ukazuje potřebu podrobnějšího rozboru, dříve než utvoříme si úsudek na závislost mezi znaky.

<sup>1)</sup> *Westergaard—Nybølle*, Grundzüge der Theorie der Statistik, 2 vyd., 1928, str. 509.



Příklad č. 27. Úmrtnost děvčat ve věku do pěti let podle původu v Dánsku v letech 1911—1915.

Věková skupina	děti manželské	děti nemanželské
0—1 měsíce	328 ‰	531 ‰
1—3 měsíce	97 ‰	221 ‰
3—6 měsíců	67 ‰	144 ‰
6—12 měsíců	42 ‰	73 ‰
1—2 roků	13 ‰	18 ‰
2—3 roků	6 ‰	5 ‰
3—4 roků	4 ‰	3 ‰
4—5 roků	3 ‰	2 ‰

Z rozdílů poměrných četností jednotlivých skupin věkových lze bezpečně souditi, že úmrtnost nemanželských dětí v nejnižších věkových stupních jest větší než úmrtnost dětí manželských; naproti tomu nelze souditi bezpečně o starších ročnících dětí, neboť v důsledku pozdějších legitimací ubývá dětí nemanželských a přibývá manželských, nikoli však naopak. Děti později legitimované a pak zemřelé připadají pak ovšem k tíži úmrtnosti dětí manželských.

Je zde tedy závislost mezi znakem illegitimity dětí a jejich úmrtností (resp. závislost nepřímá mezi legitimitou dětí a úmrtností) v nejnižších věkových skupinách; poněvadž se však soubor sám uvnitř ve vyšších věkových skupinách mění, při čemž změny ty nemůžeme zjistiti, nemáme-li pro ně též údajů o legitimacích, nelze z něho činiti závěrů platných pro celý soubor.

## ZÁVISLOST MEZI ZNAKY KVANTITATIVNÍMI (KORELACE).

### 1. Pojem korelace a srovnávání statistických řad.

Závislost mezi znaky kvantitativními, pro kterou se vžil označení korelace (též souvztažnost, kovariace) se zjišťuje a měří zvláštními metodami podle zvláštní povahy kvantitativních znaků.

Obdobně jako závislost znaků kvalitativních znamená závislost znaků kvantitativních, že s měnícími se hodnotami znaku jednoho mění se rozložení a pravděpodobnost znaku druhého. Znaky jsou přímo (kladně) korelovány, jestliže se vzrůstem hodnot jednoho znaku stávají se pravděpodobnějšími vysoké hodnoty druhého znaku, a nepřímo (záporně) korelovány, jestliže se vzrůstem hodnot jednoho znaku stávají se pravděpodobnějšími nízké hodnoty znaku druhého. Při tom opět není to závislost pevná, funkční, kde by každé hodnotě znaku jednoho odpovídala jediná určitá hodnota znaku druhého, nýbrž zvláštní závislost kolektivní, vlastní kolektivním souborům, kde každé hodnotě znaku jednoho odpovídá opět řada možných hodnot znaku druhého, a vztah mezi znaky jest opět jen pravděpodobný.

Primitivním způsobem lze zjišťovati závislost mezi znaky kvantitativními tak, že prostě sledujeme, jak se mění (rostou nebo se zmenšují) kvantitativní znaky ve dvou statistických řadách (obyčejně časových). Tento způsob nazýváme srovnáváním statistických řad. Jestliže hodnoty jednoho znaku se mění souhlasně s hodnotami znaku druhého, mluvíme o *souběžnosti* (*paralelismu*) řad, jestliže se mění opačně než četnosti znaku druhého, o *protichůdnosti* (*antagonismu*) řad; zvláště názorně se

projevují tyto pohyby při grafickém znázornění v podobě dvou čar na diagramu.

Z paralelismu nebo antagonismu řad se pak usuzuje, ovšem jen zcela zhruba, na přímou nebo nepřímou závislost znaků v řadách těch obsažených.

Takovými primitivními methodami hledána byla závislost mezi cenou žita, jako měřítko ceny nejnnutnější životní potřeby širokých vrstev obyvatelstva, a mezi úmrtností nebo kriminalitou, vyjádřenou poměrnou četností přestupků krádeže, jako měřítko kriminality těchto vrstev; vše to sledováno v řadách časových.

**Poznámka.** Srv. již historické pojednání *B. Weisz*: „Über einige wirtschaftliche und moralische Wirkungen hoher Getreidepreise,“ *Conrad's Jahrbücher*, N. F. III/1881, často citované v učebnicích statistiky.

## 2. Měření korelace korelačními tabulkami.

Korelaci mezi znaky kvantitativními můžeme zkoumati jednak (předběžně) t. zv. korelačními tabulkami, jednak vyjádřiti ji přesně pomocí některých vzorců. Korelační tabulka sestavuje se tím způsobem, že se pozorovaný soubor upravuje v tabulku současně se zřetelem k oběma znakům v něm obsaženým tak, že se jednotlivá pozorování třídí podle hodnot znaku jednoho nebo podle tříd četností tohoto znaku, a současně zařazují do tříd znaku druhého. Každá hodnota souboru je takto zařazena zároveň do příslušné třídy znaku jednoho a do příslušné třídy znaku druhého.

**Příklad č. 28.** *Korelace mezi věkem vdaných žen a počtem dětí, které se v nynějším manželství narodily (Slovensko r. 1930).*

V této korelační tabulce se objevuje rozložení charakteristické pro přímou kladnou korelaci obou znaků, neboť nejvyšší třídni četnosti sestupují s levé strany tabulky shora k pravé straně dolů, seskupeny kolem diagonály. Zároveň je to příklad, třeba ne dokonalý, t. zv. korelace lineární, t. j. takové, kde největší hustoty četností lze přibližně vyrovnati přímkou a vyjádřiti též analytickou rovnicí přímky (přímka tato se nazývá podle *Galtona* „regresní přímka“). Při korelaci záporné, nepřímé, postupují třídni četnosti vzestupně od levé strany tabulky zdola k pravé straně vzhůru. Není-li pravidelnosti v rozložení největší hustoty četností, takže je nelze vyrovnati přímkou, mluvíme o korelaci nelineární.

Při sestřovování korelačních tabulek je opět dbáti pravidla, že intervaly mezi třídami musí býti rovné.



# Korelace mezi trváním manželství vdaných žen a počtem dětí

Počet vdaných žen, kterým se v nynějším

Trvání manželství v dokončených letech	0	1	2	3	4	5	6	7	8	9
0—4	50.444	52.943	23.575	5.098	956	146	19	1	—	—
5—9	15.788	25.398	37.226	27.426	14.819	5.943	1.893	515	167	39
10—14	10.367	11.054	19.656	20.675	18.547	13.812	8.121	3.543	1.419	539
15—19	5.207	4.646	7.243	7.886	8.090	7.403	6.283	4.650	3.123	1.700
20—24	4.525	4.434	6.657	7.664	7.690	6.987	6.327	5.350	4.223	3.053
25—29	3.336	3.162	4.905	5.704	6.005	5.665	5.277	4.515	3.912	2.963
30—34	2.653	2.567	3.844	4.786	5.205	5.142	4.844	4.252	3.817	3.198
35 a více	4.404	4.083	6.010	7.340	8.322	8.328	8.466	7.751	7.202	6.014
Úhrn...	96.724	108.287	109.116	86.519	69.635	53.426	41.224	30.577	23.863	17.506

Pramen: Čs. statistika, sv. 126, str. 26. Případy, zahrnuté v rubrikách

## 3. Koefficient korelace a korelační poměr.

Nejčastěji užívaným vzorcem pro měření korelace, který se hodí ke zjištění a měření lineární korelace mezi dvěma znaky, je t. zv. *Pearsonův* koefficient korelace, označovaný písmenem  $r$ , který se zakládá na srovnávání směrodatných odchylek obou variabilních znaků. Jeho

$$\text{vzorec je: } r = \frac{\sum (x y)}{N \sigma_x \sigma_y}, \text{ ve zkrácené úpravě: } r = \frac{\sum (x y)}{\sqrt{\sum (x^2) \sum (y^2)}}$$

kde  $x$  znamená odchylky jednotlivých hodnot jednoho znaku od aritmetického průměru celé řady těchto hodnot,  $y$  odchylky hodnot druhého znaku od aritmetického průměru celé řady hodnot tohoto druhého znaku. Tento vzorec je zjednodušený tvar korelačního koeficientu, za jehož základ bereme aritmetický průměr. Původní vzorec se zakládá na odchylkách hodnot pozorovaných (empirických) od hodnot vypočtených (theoretických) na základě vyrovnaného souboru methodou nejmenších čtverců.

Koefficient korelace se pohybuje mezi 0 a 1 v případě korelace kladné a mezi 0 a —1 v případě korelace záporné.

živě narozených v tomto manželství na Slovensku r. 1930.

manželství živě narodilo dětí:

10	11	12	13	14	15	16	17	18	19	20 a více	Úhrn
—	—	—	—	—	—	—	—	—	—	—	133.176
13	2	—	—	—	—	—	—	—	—	—	129.229
237	72	47	22	5	—	2	1	—	—	—	108.059
899	365	179	75	30	12	9	—	1	—	—	57.801
1.981	994	640	241	111	54	28	11	7	5	4	60.986
2.119	1.339	775	406	194	96	51	12	12	3	4	50.456
2.270	1.490	1.002	444	272	120	48	22	15	1	3	45.995
4.791	3.130	2.387	1.110	633	288	150	47	35	13	20	80.524
12.310	7.392	5.030	2.298	1.245	570	288	93	70	22	31	666.226

označených „neudáno“, pro oba znaky, byly vypuštěny.

Ke zjištění, zda koeficient korelace, ke kterému jsme výpočtem dospěli, lze považovati za spolehlivé měřítko či zda vyjadřuje pouze nahodilé okolnosti, slouží opět srovnání s theoretickou střední odchylkou tohoto koeficientu. Theoretická střední odchylka Pearsonova koeficientu se rovná  $\sigma_r = \frac{1 - r^2}{\sqrt{N}}$ , kde  $N$  znamená počet párů pozorovaných případů.

Aby bylo možno koeficient korelace považovati za spolehlivého ukazatele závislosti dvou znaků kvantitativních, vyžaduje se obvykle, podobně jako tomu bylo u jiných poměrů, aby koeficient byl větší než trojnásobek jeho theoretické střední odchylky.

Také tehdy, jestliže zjistíme vysoký stupeň korelace (korelační koeficient je zlomek, blíží se jedničce), je třeba si uvědomiti, že ze statistické závislosti nelze ještě souditi na vztah příčinný a že závislost mezi dvěma znaky může býti zprostředkována závislostí na znaku jiném.

Jestliže je statistický soubor rozdělen podle znaku kvantitativního ve třídy, vycházíme při výpočtu korelačního koeficientu obdobně od

*Korelace mezi trváním manželství vdaných žen a počtem dětí*

Počet vdaných žen, kterým se v nynějším

Trvání manželství v dokonče- ných letech	0	1	2	3	4	5	6	7	8	9
0—4	50.444 9	52.943 6	23.575 3	5.095 0	956 3	146 6	13 9	1 12	—	—
5—9	15.788 6	25.398 4	37.226 2	27.426 0	14.819 2	5.943 4	1.893 6	515 8	167 10	39 12
10—14	10.367 3	11.054 2	19.656 1	20.615 0	18.547 1	13.812 2	8.121 3	3.543 4	1.419 5	539 6
15—19	5.207 0	4.646 0	7.243 0	7.886 0	8.090 0	7.403 0	6.283 0	4.650 0	3.123 0	1.700 0
20—24	4.625 3	4.434 2	6.667 1	7.664 0	7.690 1	6.987 2	6.327 3	5.350 4	4.223 5	3.053 6
25—29	3.336 6	3.162 4	4.905 2	5.704 0	6.006 2	5.665 4	5.277 6	4.515 8	3.912 10	2.963 12
30—34	2.653 9	2.567 6	3.844 3	4.786 0	5.205 3	5.142 6	4.844 9	4.252 12	3.817 15	3.198 18
35 a více	4.404 12	4.083 8	6.010 4	7.340 0	8.322 4	8.328 8	8.466 12	7.751 16	7.202 20	6.014 24
Úhrn...	96.724	108.287	109.116	86.519	69.635	53.426	41.224	30.577	23.863	17.506

středů tříd, jako jsme učinili při výpočtu směrodatných odchylek u takových souborů. Další postup záleží v tom, že se korelační tabulka, ve které je soubor rozložen, rozdělí na čtyři pole tím, že se ony třídy, do kterých spadá aritmetický průměr řady  $x$  a řady  $y$ , označí jako pomocný prozatímní bod, od kterého vycházíme (podobně jako tomu bylo u prozatímního aritmetického průměru). Dostaneme tím rozdělení tabulky, podle kterého vypočítáváme odchylky tříd od tříd pomocných a pak je vzájemně násobíme, abychom dostali potřebnou veličinu  $\Sigma(xy)$  pro vzorec koeficientu korelace.



živě narozených v tomto manželství.

manželství živě narodilo dětí:

10	11	12	13	14	15	16	17	18	19	20 a více	Úhrn
—	—	—	—	—	—	—	—	—	—	—	133.176
13 14	2 16	—	—	—	—	—	—	—	—	—	129.229
237 7	72 8	47 9	22 10	5 11	— 12	2 13	1 14	—	—	—	108.059
899 0	365 0	179 0	75 0	30 0	12 0	9 0	— 0	1 0	—	—	57.801
1.981 7	994 8	640 9	241 10	111 11	54 12	28 13	11 14	7 15	5 16	4 17	60.986
2.119 14	1.839 16	775 18	406 20	194 22	96 24	51 26	12 28	12 30	8 32	4 34	50.456
2.270 21	1.490 24	1.002 27	444 30	272 33	120 36	48 39	22 42	15 45	1 48	8 51	45.995
4.791 28	3.130 32	2.387 36	1.110 40	633 44	288 48	150 52	47 56	35 60	13 64	20 68	80.524
12.310	7.392	5.030	2.298	1.245	570	288	93	70	22	31	666.226

Výpočet korelačního koeficientu předvedeme na schematickém příkladě:

Příklad č. 29:

buďtež dány hodnoty znaku  $x$  1, 2, 3, 4, 5  
a hodnoty znaku  $y$  2, 5, 3, 8, 7,  
potom aritmetický průměr znaku  $x$   $A_x = 3$ ,  
aritmetický průměr znaku  $y$   $A_y = 5$ ,  
směrodatná odchylka znaku  $x$   $\sigma_x = 1,41$ ,  
směrodatná odchylka znaku  $y$   $\sigma_y = 2,28$ .

Další postup je:

hodnoty znaku		odchylky od aritme- tického průměru		dvojmoci odchylek		součin odchylek
$x$	$y$	$A_x = 3$	$A_y = 5$	$x^2$	$y^2$	$xy$
		$x$	$y$			
1	2	-2	-3	4	9	+6
2	5	-1	0	1	0	0
3	3	0	-2	0	4	0
4	8	+1	+3	1	9	+3
5	7	+2	+2	4	4	+4
				$\Sigma(x^2) = 10$	$\Sigma(y^2) = 26$	$\Sigma(xy) = 13$

$$r = \frac{13}{5 \times 1,41 \times 2,28} = 0,809.$$

Můžeme tento koeficient korelace srovnati s jeho směrodatnou odchylkou:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} = \frac{1 - 0,6545}{\sqrt{5}} = \frac{0,3455}{2,236} = 0,154.$$

$3 \sigma_r$  se rovná = 0,462, je tedy korelační koeficient větší než trojnásobná (a zřejmě i než pětinasobná) jeho směrodatná odchylka, což se podle konvence považuje za dostatečný průkaz jeho spolehlivosti.

Výpočet korelačního koeficientu v př. č. 28:

Třídy, do kterých spadá aritmetický průměr, t. j. třídu 3 u znaku  $x$  a třídu 15—19 u znaku  $y$  zvolíme za základní pomocný bod a označíme nulou 0. (To ovšem předpokládá, že dříve zjistíme aritmetický průměr obou řad,  $A_x = 3,408$  (dětí) a  $A_y = 16,115$  (let), a dále též směrodatné odchylky  $\sigma_x = 2,333$  třídních jednotek neboli dětí, ježto třídní interval pro  $x = 1$ ,  $\sigma_y = 2,394$  třídních jednotek neboli 11,97 let, ježto třídní interval pro  $y = 5$ .) Vzdálenosti  $C_x$  a  $C_y$  udávají nám vzdálenosti aritmetických průměrů od základního pomocného bodu tabulky.

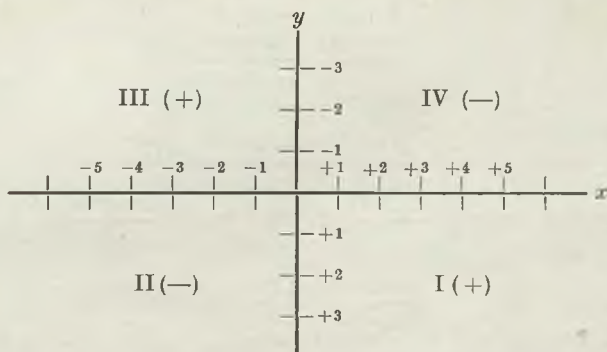
Vzdálenost  $C_x = 3 - 3,408 = -0,408$ .

Vzdálenost  $C_y = 17,5 - 16,115 = 1,385$ , obojí vzdálenosti jsou vyjádřeny v jednotkách třídních.

Součin  $C_x C_y = -0,408 \times 1,385 = 0,565$ .

Dále je třeba vypočítati  $xy$ , t. j. součin vzdáleností středů tříd od středu základní pomocné třídy 0. Tyto součiny jsou v tabulce tištěny ležatým písmem. Pro každou třídu je třeba součin ještě násobiti příslušnou četností. Při tom, je-li řada uspořádána od shora dolů vzestupně

(pro  $y$ ) a s levé strany do pravé rovněž vzestupně (pro  $x$ ), jak je zvykem, dostaneme čtyři pole, rozdělená oběma osami kolmými, které tvoří třídy, jež jsme učinili pomocnými body, a označili jako 0, podle schematu:



Pole druhé a čtvrté má četnosti  $xy$  záporné, poněvadž jeden ze součinitelů je kladný a druhý záporný. Pole první a třetí má četnosti  $xy$  kladné, poněvadž oba součinitelé jsou kladní nebo oba záporní.

Potom je možno přikročiti k vypočítání hodnoty  $\frac{\sum (x y f)}{N}$  podle jednotlivých polí.

Uvádíme podrobně výpočet třetího (kladného) a druhého (záporného) pole, výpočet prvního a čtvrtého pole pro jejich obsáhlost jen úhrnně.

III. pole:

$$\begin{array}{rcl}
 19.656 \times 1 & = & 19.656 \\
 11.054 \times 2 & = & 22.108 \\
 10.367 \times 3 & = & 31.101 \\
 37.226 \times 2 & = & 74.452 \\
 25.398 \times 4 & = & 101.592 \\
 15.788 \times 6 & = & 94.728 \\
 23.575 \times 3 & = & 70.725 \\
 52.943 \times 6 & = & 317.658 \\
 50.444 \times 9 & = & 453.996 \\
 \hline
 \text{úhrn} & = & 1,186.016
 \end{array}$$

II. pole:

$$\begin{array}{rcl}
 6.657 \times 1 & = & 6.657 \\
 4.434 \times 2 & = & 8.868 \\
 4.525 \times 3 & = & 13.575 \\
 4.905 \times 2 & = & 9.810 \\
 3.162 \times 4 & = & 12.648 \\
 3.336 \times 6 & = & 20.016 \\
 3.884 \times 3 & = & 11.532 \\
 2.567 \times 6 & = & 15.402 \\
 2.653 \times 9 & = & 23.877 \\
 6.010 \times 4 & = & 24.040 \\
 4.083 \times 8 & = & 32.664 \\
 4.404 \times 12 & = & 52.848 \\
 \hline
 \text{úhrn} & = & 231.937
 \end{array}$$

úhrn čtvrtého pole činí 173.121, úhrn prvního pole činí 1,825.237.



$$\text{Summa } \Sigma(xyf) = \begin{array}{r} \text{I. pole } +1,825.237 \\ \text{III. pole } +1,186.016 \\ \hline +3,011.253 \\ \text{II. pole } -231.937 \\ \text{IV. pole } -173.121 \\ \hline -405.058 \\ \text{úhrn } +2,606.195 \end{array}$$

$$\frac{\Sigma(xyf)}{N} = \frac{2,606.195}{666.226} = 3,912$$

$$\text{Skutečná hodnota } \frac{\Sigma(xyf)}{N} \text{ se pak rovná } = \frac{\Sigma(xyf)}{N} - C_x C_y = \\ = 3,912 + 0,565 = 4,477,$$

$$r = \frac{\Sigma(xyf)}{N \sigma_x \sigma_y} = \frac{4,477}{5,585} = 0,801.$$

Směrodatná odchylka tohoto korelačního koeficientu se rovná

$$\sigma_r = \frac{1-r^2}{\sqrt{N}} = \frac{1-0,801^2}{\sqrt{666.226}} = \frac{1-0,6416}{816} = 0,00043.$$

$3 \sigma_r = 0,00129 < 0,801$ , korelační koeficient lze tedy považovati za spolehlivou míru korelace.

Nelineární korelací nazýváme takový případ závislosti mezi znaky kvantitativními, při které vzájemné uspořádání souboru podle obou znaků v korelační tabulce nelze vyrovnati přímkou, nýbrž jinou čarou, obyčejně křivkou. Nejjednodušší případ nelineární korelace je vyrovnání křivkou druhého stupně podle metody nejmenších čtverců. Ke zjištění nelineární korelace hodí se spíše výpočet korelačního poměru, míry to korelace, kterou srovnáváme s koeficientem korelace, abychom zjistili velikost odchylky dané korelace od korelace lineární.

Vzorec korelačního poměru je  $\eta = \frac{\sigma_{Ay}}{\sigma_y}$ , kde  $\sigma_{Ay}$  znamená směrodatnou odchylku průměrů hodnot  $y$  v každé třídě znaku  $x$  od aritmetického průměru všech hodnot znaku  $y$ , a  $\sigma_y$  značí opět směrodatnou odchylku jednotlivých hodnot  $y$  od aritmetického průměru hodnot  $y$ . Je tedy třeba vypočítati nejprve pro každou hodnotu  $x$  (resp. pro každou třídu znaku  $x$ ) průměr hodnoty znaku  $y$  na ni připadající, potom vypočísti rozdíl této hodnoty od aritmetického průměru znaku  $y$  (pro celý soubor). Tento rozdíl se pak umocní dvěma (a násobí četností příslušné třídy znaku), pak dělí počtem všech pozorování a odmocňuje dvěma, stejně jako při výpočtu směrodatné odchylky.

Příklad č. 30. Korelace mezi výškou těla otců a synů. (Yule Úvod do theorie statistiky, č. překlad, str. 161.)

Výška těla synů v palcích	Výška těla otců v palcích																	Úhrn
	58,5	59,5	60,5	61,5	62,5	63,5	64,5	65,5	66,5	67,5	68,5	69,5	70,5	71,5	72,5	73,5	74,5	
	59,5	60,5	61,5	62,5	63,5	64,5	65,5	66,5	67,5	68,5	69,5	70,5	71,5	72,5	73,5	74,5	75,5	
59,5–60,5	—	—	—	—	0,5	0,5	1	—	—	—	—	—	—	—	—	—	—	2
60,5–61,5	—	—	—	—	0,5	—	—	—	1	—	—	—	—	—	—	—	—	1,5
61,5–62,5	—	0,25	0,25	—	0,5	1	0,25	0,25	0,5	0,5	—	—	—	—	—	—	—	3,5
62,5–63,5	—	0,25	0,25	2,25	2,25	2	4	5	2,75	1,25	—	0,25	0,25	—	—	—	—	20,5
63,5–64,5	1	—	1,5	3,75	3	4,25	8	9,25	3	1,25	1,5	0,75	1,25	—	—	—	—	38,5
64,5–65,5	2	1	0,5	2	3,25	9,5	13,5	10,75	7,5	5,5	3,5	2,5	—	—	—	—	—	61,5
65,5–66,5	—	0,5	1	2,25	5,25	9,5	10	16,75	17,5	16	5,25	2	2,5	1	—	—	—	89,5
66,5–67,5	—	1,5	2	4,75	3,5	13,75	19,75	26,5	25,75	19,5	12,5	13,75	3,25	0,5	1	—	—	148
67,5–68,5	—	—	1,5	2	7,5	10	10,25	24,25	31,5	23,5	29,5	13,25	8,5	9,5	2,25	—	—	173,5
68,5–69,5	—	—	1	—	5,25	5	12,75	18,25	16	24	29	21,5	10	3,5	2,25	—	1	149,5
69,5–70,5	—	—	—	—	1	2,5	5,75	18,75	11,75	19,5	22,5	19,5	14,5	6,25	3,5	1,5	1	128
70,5–71,5	—	—	—	—	—	3,25	5	8,75	10,75	19	14,75	20,75	10,75	8	5	1	1	108
71,5–72,5	—	—	—	—	—	0,25	3	1,25	7	7,75	10,75	11,25	10	8,5	2,75	0,5	—	63
72,5–73,5	—	—	—	—	—	—	0,75	0,75	2,5	7,5	6,5	6	7,5	6,25	3,25	0,5	0,5	42
73,5–74,5	—	—	—	—	1	—	1,5	1,5	—	5,25	2,25	2,5	6,5	3,25	3,25	—	2	29
74,5–75,5	—	—	—	—	—	—	—	—	—	1	2	—	2,5	0,75	1,75	0,5	—	8,5
75,5–76,5	—	—	—	—	—	—	—	—	—	1,25	0,25	—	0,5	1	1	—	—	4
76,5–77,5	—	—	—	—	—	—	—	—	—	1,25	0,25	1	—	—	1,5	—	—	4
77,5–78,5	—	—	—	—	—	—	—	—	—	—	1	1	—	0,25	0,75	—	—	3
78,5–79,5	—	—	—	—	—	—	—	—	—	—	—	—	—	0,25	0,25	—	—	0,5
Úhrn...	3	3,5	8	17	33,5	61,5	95,5	142	137,5	154	141,5	116	78	49	28,5	4	5,5	1078

Vzhledem k tomu, že se korelační poměr zakládá na dvojmocích odchylek, je vždy číslem kladným, a dá se dále odvoditi, že nemůže býti menší než korelační koeficient  $r$  pro tytéž páry znaků  $x$  a  $y$ .

**Příklad č. 30.** (Yule, Úvod do theorie statistiky, český překlad, str. 161.)

Jednotlivé charakteristiky tohoto souboru jsou:

Aritmetický průměr hodnot  $x$   $A_x = 67,70$  palců

Aritmetický průměr hodnot  $y$   $A_y = 68,66$  „

Směrodatná odchylka pro  $x$   $\sigma_x = 2,72$  „

Směrodatná odchylka pro  $y$   $\sigma_y = 2,75$  „

Míra korelace  $r = +0,51$  „

*Výpočet korelačního poměru mezi výškou těla synů a výškou těla otců.*  
(Yule, l. c. str. 209.)

Střed tříd znaku $x$ (výška těla otců)	Aritmetické průměry sloupců zna- ku $y$ (výška těla synů)	Rozdíl od aritm. průmě- ru výšky těla všech synů ( $A_y = 68,66$ )	Čtverec tohoto rozdílu	Četnost jednotlivých sloupců	Součin četnosti a dvojmočí rozdílu
59	64,67	-3,99	15,9201	3	47,76
60	65,64	-3,02	9,1204	3,5	31,92
61	66,34	-2,32	5,3824	8	43,06
62	65,66	-3,10	9,6100	17	163,37
63	66,68	-1,98	3,9204	33,5	131,33
64	66,74	-1,92	3,6864	61,5	226,71
65	67,19	-1,47	2,1609	95,5	206,37
66	67,61	-1,05	1,1025	142	156,56
67	67,95	-0,71	0,5041	137,5	69,31
68	69,07	+0,41	0,1681	154	25,89
69	69,39	+0,73	0,5329	141,5	75,41
70	69,74	+1,08	1,1664	116	135,30
71	70,50	+1,84	3,3856	78	264,08
72	70,87	+2,21	4,8841	49	239,32
73	72,-	+3,34	11,1556	28,5	317,93
74	71,60	+2,84	8,0656	4	32,26
75	71,73	+3,07	9,4249	5,5	51,84
Součet				1.078	2.218,42

$$\sigma^2_{Ay} = 2.218,42 : 1.078 = 2,058, \sigma_{Ay} = \sqrt{2,058} = 1,43, \eta = \frac{\sigma_{Ay}}{\sigma_y} = \frac{1,43}{2,75} = 0,52.$$



Postup pro výpočet hodnoty znaku  $y$  pro každou třídu znaku  $x$  je patrný z výpočtu pro třídu první a druhou znaku  $x$ :

$$\begin{array}{rcl} x = 59 & 64 \times 1 & = 64 \\ & 65 \times 2 & = 130 \\ & \hline & 194 : 3 & = 64,67 \end{array}$$

$$\begin{array}{rcl} x = 60 & 62 \times 0,25 & = 15,50 \\ & 63 \times 0,25 & = 15,75 \\ & 65 \times 1 & = 65 \\ & 66 \times 0,5 & = 33 \\ & 67 \times 1,5 & = 100,50 \\ & \hline & 229,75 : 3,5 & = 65,64 \end{array}$$

t. j. střed každého intervalu třídního násobíme četností třídy znaku  $y$ , úhrn součinů pak dělíme počtem všech případů.

Poněvadž je korelační poměr vždy číslem kladným, usuzujeme na kladnou nebo zápornou korelaci podle podoby korelační tabulky.

Je-li korelační poměr značně vyšší než korelační koeficient, jde o korelaci nelineární. V daném příkladě rozdíl je malý

$$\eta - r = 0,52 - 0,51 = 0,01, \text{ je tedy korelace lineární.}$$

### 3. Korelace dílčí.

Jestliže máme danu závislost mezi více než dvěma znaky kvantitativními, nebo domníváme-li se, že závislost mezi dvěma znaky kvantitativními je zprostředkována jejich závislostí na znaku třetím (nebo na více znacích jiných), mluvíme o mnohonásobné korelaci. Hledáme-li v souboru rozděleném na př. podle tří znaků korelaci mezi dvěma znaky pro určitou danou hodnotu znaku třetího, nazývá se tato korelace dílčí (parciální), neboť je určena koeficientem dílčí korelace, označovaným  $r_{12.3}$  v té části souboru, v níž má první znak jistou danou (konstantní) hodnotu. (srv. Janko Základy statistické indukce, str. 79, tam též další postup při výpočtu dílčích korelačních koeficientů).

### 4. Korelace u řad časových.

Poněvadž v řadách časových vznikají často klamné nebo jen zdánlivé korelace mezi kvantitativními znaky, používá se při výpočtu korelace metody zvláštní, že se nesrovnávají hodnoty obou znaků, nýbrž jejich změny v časových obdobích. Postupuje se tak, že se odečítá vždy hodnota následující od hodnoty předcházející a z těchto rozdílů se teprve počítá korelace stejnou metodou, jako u jiných souborů. Tím se snažíme odstraniti vliv náhodných stejně u obou znaků působících

příčin, které v časovém vývoji, a tedy i v časových řadách, často se vyskytují a vyvolávají zdánlivé korelace různého stupně. Někdy ani to se ještě nepovažuje za dostatečné, a počítají se druhé rozdíly t. j. rozdíly ze dvou sousedních rozdílů hodnot znaků. Podobně je možno postoupiti ještě k vyšším rozdílům, třetím nebo čtvrtým, a z nich teprve počítati korelační koeficient.

**Příklad č. 31.** *Korelace mezi natalitou a úmrtností v Čechách v letech 1893 až 1913.*

Rok	Na 1000 obyvatel připadalo		Rozdíly od aritmetického průměru		Dvojmoči odchylek		Součiny	
	živě narozených	zemřelých	$A_x=32.9$	$A_y=22.7$	$x^2$	$y^2$	$+xy$	$-xy$
1893	36.9	27.5	+4.0	+4.8	16.00	23.04	19.20	
1894	36.4	26.5	+3.5	+3.8	12.25	14.44	13.30	
1895	37.1	25.3	+4.2	+2.6	17.64	6.76	10.92	
1896	37.3	24.4	+4.4	+1.7	19.36	2.89	7.48	
1897	36.8	24.5	+3.9	+1.8	15.21	3.24	7.02	
1898	36.6	24.2	+3.7	+1.5	13.69	2.25	5.55	
1899	36.7	25.3	+3.8	+2.6	14.44	6.76	9.88	
1900	34.4	24.2	+1.5	+1.5	2.25	2.25	2.25	
1901	34.1	23.3	+1.2	+0.6	1.44	0.36	0.72	
1902	34.8	23.4	+1.9	+0.7	3.61	0.49	1.33	
1903	33.0	22.9	+0.1	+0.2	0.01	0.04	0.02	
1904	33.2	22.7	+0.3	0	0.09	—	—	
1905	30.9	23.8	—2.0	+1.1	4.00	1.21		—2.2
1906	32.5	20.3	—0.4	—2.4	0.16	5.76	0.96	
1907	31.6	20.9	—1.3	—1.8	1.69	3.24	2.34	
1908	31.8	20.9	—1.1	—1.8	1.21	3.24	1.98	
1909	29.7	20.6	—3.2	—2.1	10.24	4.41	6.72	
1910	28.5	19.4	—4.4	—3.3	19.36	10.89	14.52	
1911	27.6	19.9	—5.3	—2.8	28.09	7.84	14.84	
1912	26.3	19.7	—6.6	—3.0	43.56	9.00	19.80	
1913	25.8	18.1	—7.1	—4.6	50.41	21.16	32.66	
					274.71	129.27	167.44	—2.2

$$A_x = 32,9, A_y = 22,7$$

$$\Sigma (x^2) = 274,71$$

$$\Sigma (y)^2 = 129,27$$

$$\Sigma (xy) = 165,24$$

$$r = \frac{\Sigma (x y)}{\sqrt{\Sigma (x)^2 \Sigma (y)^2}} = 0,876$$

Střední theoretická odchylka tohoto korelačního koeficientu

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} = \frac{1 - 0,767}{\sqrt{21}} = 0,051$$

Trojnásobná střední odchylka  $3 \sigma_r = 0,153$ , korelační koeficient převyšuje tedy značně trojnásobek střední odchylky.

Počítáme-li však korelaci nikoli přímo z hodnot (poměrných četností) natality z úmrtnosti, nýbrž z rozdílů prvního stupně z obou řad, obdržíme tabulku:

**Pokračování příkl. č. 31. Korelace mezi natalitou a úmrtností v Čechách.**

Rozdíly prvního stupně		Odchylky rozdílů od aritm. průměru		Dvojmocí odchylek		Součiny	
v řadě natality	v řadě úmrtnosti	$A_x = 0,6$	$A_y = 0,5$	$x^2$	$y^2$	$+xy$	$-xy$
+0,5	+1,0	-0,1	+0,5	0,01	0,25		-0,05
-0,7	+1,2	-1,3	+0,7	1,69	0,49		-0,91
-0,2	+0,9	-0,8	+0,4	0,64	0,16	+0,32	
+0,5	-0,1	-0,1	-0,6	0,01	0,36	+0,06	
+0,2	+0,3	-0,4	-0,2	0,16	0,04	+0,08	
-0,1	-1,1	-0,7	-1,6	0,49	2,56	+1,12	
+2,3	+1,1	+1,7	+0,6	2,89	0,36	+1,02	
+0,3	+0,9	-0,3	+0,4	0,09	0,16		-0,12
-0,7	-0,1	-1,3	-0,6	1,69	0,36	+0,78	
+1,8	+0,5	+1,2	0	1,44	0	-	-
-0,2	+0,2	-0,8	-0,3	0,64	0,09	+0,24	
+2,3	-1,1	+1,7	-1,6	2,89	2,56		-2,72
-1,6	+3,5	-2,2	+3,0	4,84	9,00		-6,60
+0,9	-0,6	+0,3	-1,1	0,09	1,21		-0,33
-0,2	-	-0,8	-0,5	0,64	0,25	+0,40	
+2,1	+0,3	+1,5	-0,2	2,25	0,04		-0,30
+1,2	+1,2	+0,6	+0,7	0,36	0,49	+0,42	
+0,9	-0,5	+0,3	-1,0	0,09	1,00		-0,30
+1,3	+0,2	+0,7	-0,3	0,49	0,09		-0,21
+0,5	+1,6	-0,1	+1,1	0,01	1,21		-0,11
				$\Sigma(x)^2 = 21,41$	$\Sigma(y)^2 = 20,68$	$\Sigma(xy) = 4,44$	$\Sigma(xy) = -11,65$

$$A_x = +0,6, A_y = +0,5$$

$$\Sigma(x)^2 = 21,41$$

$$\Sigma(y)^2 = 20,68$$

$$\Sigma(xy) = -7,21$$

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma(x)^2 \Sigma(y)^2}} = -0,342$$



Prostým srovnáním řad natality a úmrtnosti za dvacet jeden rok v Čechách pozorujeme velkou souběžnost obou řad, které obě vykazují výraznou sestupnou tendenci. Také korelační koeficient, ke kterému jsme dospěli, je velmi vysoký a pozitivní.

Vypočteme-li však rozdíly prvního stupně v obou řadách a z nich korelační koeficient, jak se stalo v pokračování příkladu č. 31, obdržíme koeficient korelace nízký a k tomu negativní. Je patrné, že závislost mezi oběma soubory je spíše vyvolána společnými vzdálenými příčinami (mluvívá se o sekulárních vývojových tendencích), než jejich bezprostřední spojitostí. Korelaci v řadách časových je potřeba posuzovati zvláště opatrně.

**Literatura ke kapitole 10—11.** Nejvýznamnější spisy citovány jsou v textu.

## NÁHODNÉ VÝBĚRY.

Již v kapitole 4, odd. II. jsme srovnávali všeobecné theoretické rozložení souboru s jeho rozložením skutečným (empirickým), na př. výsledky vrhů mincí s theoretickou (mathematickou) jejich pravděpodobností. Nyní jde o speciálnější případ a problém, totiž srovnávati theoretické rozložení četností, vypočítané podle předpokládané hypotézy se skutečným rozložením četností pozorovaného souboru. Pozorovaný soubor považujeme při tom za výběr, vybraný vzorek, z theoretického souboru.

Problém záleží prakticky v tom, že při statistických šetřeních velmi často není možno zachytiti vyčerpávajícím způsobem celý soubor, nýbrž že se musíme spokojiti pouze se šetřením, kterým zachycujeme toliko výběr z celkového souboru. Je ovšem otázka, můžeme-li považovati výsledky statistického popisu získané na tomto výběru nebo vzorku za platné též pro celý soubor, tedy na př. můžeme-li považovati aritmetický průměr získaný z pozorovaného výběru za charakteristický též pro celý soubor.

*Janko* (Základy statistické indukce, str. 95) nazývá zobecnění statistických výsledků, získaných zpracováním určitého souboru, za předpokladu, že je možno vztahovati hodnoty zjištěných charakteristik na rozsáhlejší soubor než je ten, ze kterého byly skutečně odvozeny, *statistickou indukci*.

Zkoumati, zda můžeme považovati výsledky na vybraném vzorku za platné pro celý soubor, umožňuje t. zv. *test Pearsonův*, označovaný řeckým písmenem  $\chi^2$ . Test tento se zakládá na srovnání hodnot theoretických a skutečných v pozorovaném výběru a rovná se součtu dvoj-

močí těchto rozdílů uvedených v poměr k theoretické četnosti výběru

$$X^2 = \sum \left( \frac{(\overline{m}_r - m_r)^2}{m_r} \right), \text{ kde } \overline{m} \text{ značí nám hodnoty skutečné (pozo-}$$

rované), v  $r$ -tém výběru,  $m$  hodnoty theoretické v témže výběru,  $m_r$  theoretickou četnost v témž výběru.

Z tohoto vzorce je patrné, že rozdíly mezi hodnotami theoretickými a empirickými jsou ve čtverci a tedy vždy kladné, a že čím jsou jednotlivé ty rozdíly větší, tím je větší i celkový součet a tudíž i  $X^2$ . V případě, že by obě rozložení, theoretické i skutečné, byla shodná, součet rozdílů a tedy i  $X^2$  by se rovnalo nule.

Theoretickou pravděpodobnost a theoretické rozložení souborů vzniklých z vrhu mincí a jiných her náhodných, kde podmínky očekávané události jsou známé, lze vypočítati podle počtu pravděpodobností a srovnávati pak s výsledky pozorovaných událostí (na př. ve výběru 1000 hodů mincí). U skutečných, empirických souborů statistických ovšem neznáme předem theoretické rozložení. Přes to můžeme použití i u těchto souborů uvedeného postupu a Pearsonova testu za určitých podmínek.

Máme-li nějakou hypotézu, která nám dává souhrnně toliko čtyři možné výsledky, pouze tři z nich jsou neodvislé, čtvrtý je dán jako nutný výsledek prvních tří. Mluvíme pak o třech stupních volnosti, o jednom méně než je počet všech případů možných.

Takové případy vyskytly se nám v kapitole o závislosti znaků kvalitativních. Zkoumali jsme na příklad závislost mezi pohlavím mužským a úmrtností v souboru obyvatelstva Čech (př. č. 24).

Předpokládejme nyní, že mezi úmrtností a pohlavím není žádné závislosti, nýbrž že znaky ty jsou nezávislé. Potom platí známá rovnice

$$(v. \text{ kap. } 10). \quad (AB) = \frac{(A)(B)}{N}.$$

Z této rovnice, která by musila platiti, kdyby oba znaky  $A$  i  $B$  byly na sobě nezávislé, jak předpokládáme, můžeme si postupně vypočítati všechny hodnoty theoretické, pro  $(AB)$ ,  $(A\beta)$ ,  $(\alpha B)$ ,  $(\alpha\beta)$ , které potřebujeme. Výsledek je patrný z tabulky dále uvedené, kde nejprve jsou vždy uvedeny hodnoty skutečné, pod nimi hodnoty theoretické.

**Příklad č. 32.** Postup je takový:

$$(AB) = \frac{(A)(B)}{N} = \frac{3,452.123 \times 95.597}{7,109.376} = 46.419,3.$$



Dále dosazujeme:

$$(A\beta) = (A) - (AB) = 3,405.703,7$$

$$(\alpha B) = (B) - (AB) = 49.177,7$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = 3,608.075,3$$

$(AB)$ (zemřelí muži)	$(A\beta)$ (muži, kteří nezemřeli)	$(A)$ (celý soubor mužů)
skutečné 49.398	3,402.725	
theoretické 46.419,3	3,405.703,7	3,452.123
$(\alpha B)$ (zemřelé ženy)	$(\alpha\beta)$ (ženy, které nezemřely)	$(\alpha)$ (celý soubor žen)
skutečné 46.199	3,611.054	
theoretické 49.177,7	3,608.075,3	3,657.253
$(B)$ (všech zemřelých)	$(\beta)$ (všichni obyvatelé, kteří nezemřeli)	$N$ (celý soubor obyvatel- stva)
95.597	7,013.779	7,109.376

Z této tabulky zjistíme rozdíly mezi hodnotami skutečnými a theoretickými a vypočteme

$$\bar{m}_1 - m_1 = 49.398 - 46.419,3 = 2.978,7$$

$$\bar{m}_2 - m_2 = 3,402.725 - 3,405.703,7 = -2.978,7$$

$$\bar{m}_3 - m_3 = 46.199 - 49.177,7 = -2.978,7$$

$$\bar{m}_4 - m_4 = 3,611.054 - 3,608.075 = 2.978,7$$

$$\begin{aligned} \text{Potom se rovná } X^2 &= \frac{2.978,7}{46.419,3} + \frac{2.978,7}{3,405.703,7} + \frac{2.978,7}{49.177,7} + \\ &+ \frac{2.978,7}{3,608.075} = 376,6 \end{aligned}$$

Pro různé hodnoty integrálu funkce  $X^2$  sestavil *Pearson* a později *Fisher* tabulku, ve které je možno naléztí jim odpovídající hodnoty pravděpodobnosti  $P$ . Za statisticky významné je zvykem považovati pouze hodnoty větší, než pro jaké je pravděpodobnost 0,05.

Tabulky tyto jsou otištěny v celém rozsahu v *Pearsonových Tables for Statisticians and Biometricians*, Londýn 1914 a 1931.

V příkladě č. 32 je hodnota  $X^2$  velmi vysoká, což předem ukazuje, že je velký stupeň asociace obou znaků, a že  $P$ , pravděpodobnost jejich ne-odvislosti, je velmi malá. Tak vysoké hodnoty  $X^2$  nejsou již obsaženy v uvedených *Pearsonových* tabulkách a je třeba k výpočtu příslušné

hodnoty  $P$  použití metody uvedené v tabulce XVII a v úvodu k Pearsonovým tabulkám na str. XIII a XXXV.

**Poznámka.** Postup je následující:

V tabulce XVII najdeme hodnoty nejbližší  $\chi^2 = 376,6$

pro 350	74,826	diference 1. řádu	diference 2. řádu
pro 400	85,655	10,829	0,003
pro 450	96,487	10,832	

$$(-\log P) = 74,826 + \frac{26,6}{50} (10,829) - \frac{1}{2} \left( \frac{26,6}{50} \right) \left( \frac{23,4}{50} \right) (0,003) = 80,587$$

$$\log P = 81,413$$

$P = \frac{1,589}{10^{81}}$ , tedy hodnota krajně nepatrná vzhledem k velikosti jmenovatele.

Na srovnání rozložení theoretického s rozložením skutečným uvnitř vybraného vzorku zakládá se rovněž kritérium *Lexisova*, zvané divergenčním koeficientem a označované písmenem  $Q$ . Kritérium to srovnává theoretickou odchylku se směrodatnou odchylkou pozorovaného souboru:

$$Q = \frac{\sqrt{\frac{\sum (x)^2}{r-1}}}{\sqrt{\frac{p \cdot q}{s}}}$$

při čemž v čitateli písmenem  $r$  označujeme počet případů ve vybraném vzorku, písmenem  $s$  ve jmenovateli průměr počtu všech případů.

Ve většině případů souborů jevů sociálních bychom nemohli použití tohoto vzorce, poněvadž neznáme theoretické pravděpodobnosti výskytu těchto jevů. V řadách časových můžeme však nahraditi pravděpodobnost theoretickou pravděpodobností empirickou t. j. bēfeme za theoretické takové poměrné četnosti, jaké jsme při pozorování velkého počtu případů zjistili. Na př. nevíme, jaká by měla býti theoretická pravděpodobnost výskytu mužského pohlaví u novorozených dětí, zda poměr 1:1 nebo 1:2 nebo nějaký jiný. Na základě pozorování velkého počtu novorozenců během dlouhé doby zjišťujeme průměrný poměr 51:100 novorozených pohlaví mužského ke všem novorozeným. Poměrnou četností pohlaví mužského mezi novorozenými tedy nahrazujeme theoretickou pravděpodobnost  $p$ , kterou neznáme.

Označíme-li tuto empirickou pravděpodobnost poměrnou četností  $\frac{m}{n}$  a nahradíme jí theoretickou pravděpodobnost  $p$  a podobně též pravdě-

podobnost  $q$ , která, jak víme, se rovná  $q = 1 - p$ , nahradíme poměrem

$1 - \frac{m}{n}$ , obdržíme divergenční koeficient v podobě

$$Q = \frac{\sqrt{\frac{\sum (x_i^2)}{r-1}}}{\sqrt{\frac{\frac{m}{n} \left(1 - \frac{m}{n}\right)}{n}}}$$

Podle toho, je-li theoretická odchylka (takto supovaná) větší nebo menší než skutečná směrodatná odchylka, a divergenční koeficient je tudíž menší nebo větší než jedna celá, mluví se o nadnormální nebo subnormální dispersi. Ve statistických souborech z oboru jevů sociálních je pravidlem nadnormální disperse.

Lexisův divergenční koeficient lze snadno převést na Pearsonův test  $X^2$ , neboť  $Q = \frac{X^2}{r-1}$ , kde  $r$  značí počet prvků, z nichž se výběr skládá.

### Příklad č. 33.

*Podíl mrtvých narozených dětí na všech porodech v Československu v letech 1925—1935.*

(Pramen: Statistická příručka RČS. IV, Statistická ročenka 1934, 1935, 1936, 1937, 1938.)

Rok	všech rozených	mrtvých rozených	ze všech roze- ných bylo mrtvých roze- ných	odchylky od průměru $\bar{x}$	dvojmoci odchylek $x^2$
1925	364.326	8.337	0,0229	0,0005	0,00000025
1926	359.955	8.249	0,0229	0,0005	0,00000025
1927	343.372	7.663	0,0223	- 0,0001	0,00000001
1928	344.794	7.525	0,0218	- 0,0006	0,00000036
1929	333.589	7.282	0,0218	- 0,0006	0,00000036
1930	340.704	7.461	0,0219	- 0,0005	0,00000025
1931	325.430	7.167	0,0215	- 0,0009	0,00000081
1932	319.761	7.304	0,0220	- 0,0004	0,00000016
1933	294.473	6.755	0,0224	0	0
1934	287.227	6.470	0,0225	0,0001	0,00000001
1935	277.008	6.083	0,0220	- 0,0004	0,00000016
Uhrn	3,590.639	80.286			0,00000262



Průměr za jedenáct let:

(zaokrouhleno

326.422

7299

0,02236

na 0,0224)

$$Q = \frac{\sqrt{\frac{\sum (x)^2}{s-1}}}{\sqrt{\frac{pq}{n}}} = \frac{\sqrt{\frac{0.000000262}{10}}}{\sqrt{\frac{0,0224 \times 0,9776}{326,422}}} = \frac{0,00051186}{0,00025884} = 1,977.$$

Poněvadž Pearsonův test se rovná  $\chi^2 = Q(r-1) = 19.77$ , kde  $r-1 = 10$  stupňů volnosti.

Tomu odpovídá v tabulkách hodnot  $\chi^2$  plocha mezi 0,05 a 0,02. Jako tedy divergenční koeficient Lexisův je dosti značně vyšší jedničky, i test Pearsonův ukazuje, že pravděpodobnost stálého poměru mezi živě a mrtvě rozenými dětmi je příliš malá, takže musíme považovati poměr tento ve vybraném časovém vzorku za náhodný.

K Pearsonovu indexu významnosti můžeme však dospěti též přímo tak, že si sestojíme theoretické rozložení souboru (základní soubor) a zjišťujeme odchylky od rozložení pozorovaného (vybraného vzorku). Chybící hypotézu o rozložení základního souboru nahrazujeme opět poměrem zjištěným za delší časové období, tedy v našem případě poměrem mrtvě narozených dětí ke všem narozeným v průměru uvedených jedenácti let.

#### Příklad č. 34.

*Poměr narozených nemanželských dětí ke všem narozeným dětem. Je stálý nebo náhodný?*

V průměru patnáctiletí 1921—1935 byl poměr dětí nemanželských ze všech (živě i mrtvě) rozených dětí roven  $m = 0,106$ . V některých letech lze nalézt velkou shodu mezi předpokládaným theoretickým rozložením a skutečným rozložením manželských a nemanželských porodů.

R. 1929.

	Počet dětí manželských	počet dětí nemanželských	Úhrn
hodnoty skutečné ( $\bar{m}$ ) . . .	298.016	35.573	333.589
hodnoty očekávané ( $m$ ). . .	298.239	35.350	333.589
$\frac{(\bar{m} - m)^2}{m}$ . . . . .	0,167	1,407	1,574

Hodnotě  $X^2 = 1,574$  odpovídá v Pearsonově tabulce pro jeden stupeň volnosti,  $r-1 = 1$ , plocha mezi 0,20 a 0,30. Vybraný vzorek za

r. 1929 tedy plně potvrzuje hypotézu očekávaného poměru nemanželských dětí (nebylo by tomu asi tak u některých jiných roků.)

Uvedené metody, podobně jako Lexisova divergenčního koeficientu, můžeme použít pouze u řad časových. Věcné řady statistické, upravené podle znaku kvantitativního, lze však srovnávat s teoretickým rozložením podle *Laplace-ova (Gaussova)* zákona, a to opět v některých případech i v oboru jevů sociálních, na př. demografických, a to použitím metody *Sheppardovy* a jeho tabulek *Laplace-ova* integrálu, obsažených v Pearsonových *Tables for Statisticians and Biometricians*.

Velká stabilita blížící se normální, podle Lexisova kriteria, vyskytuje se často též u souborů jevů, které jsou velmi řídké, tedy jejichž pravděpodobnost ve velkém souboru je velmi nepatrná, jako jsou na př. sebevraždy dětí, vzácně se vyskytující nemoci a pod. Tuto pravidelnost v časových řadách u jevů zřídka pozorovaných a majících tedy jen malou pravděpodobnost, pravidelnost, která na první pohled překvapuje, nazval *Bortkiewicz* (*Bortkiewicz Das Gesetz der kleinen Zahlen*, Lipsko 1898) zákonem malých čísel.

Normální nebo skoro normální stabilita některých jevů, které podle obecného přesvědčení a subjektivního cítění lidského jsou závislé na jeho svobodném rozhodování, vedla některé statistiky k tomu, že se snažili objasniti statistickou methodou problém svobodné vůle a vytvořili tak zajímavou kapitolu historie novodobé statistiky.

Byl to zejména *Quêtelet* a jeho žáci, kteří se snažili ve společenských jevech nalézt pevné zákony jako obdobu zákonů přírodních a kteří považovali též stabilitu časových řad, vyjadřujících jevy sociální, za výsledek působení pevných sil a tudíž za odporující učení o svobodě lidské vůle. Největším dojmem působily časové řady t. zv. morální statistiky, na př. počet sebevražd, zločinů, kde událost sama je výsledkem rozhodování a lidské úvahy. Stabilitu těchto řad vykládal *Quêtelet* a jeho škola tím, že existují jisté stálé příčiny, které nutně vedou k takovým činům, působíce neodolatelnou silou na určitá individua. Toto mechanistické pojetí takových činů, které je vykládalo obdobným způsobem, jako pohyby fysických těles nebo strojů působením fysických sil, vedlo u některých statistiků k závěru o úplném determinismu všeho lidského konání a k popření svobodné lidské vůle vůbec. *Quêtelet* sám se domníval, že společnost lidská má v sobě zárodky všech zločinů, které mají být spáchány, že tyto zločiny připravuje a že zločinec je pouhým nástrojem, který je provádí, a že počet sebevražd, vražd a zločinů je každého roku tak pevně předem stanoven, jako rozpočet státních příjmů a vydání. Od

tohoto společenského determinismu vedl pak vývoj ke společenskému fatalismu, jaký se projevuje na př. u *Adolfa Wagnera*, *Comtea*, *Bucklea*.<sup>1)</sup> V beletrii se projevuje zejména ve směru zvaném naturalismus v druhé polovině 19. stol. (zvláště ve Francii *Emile Zola* ve svém románovém cyklu o rodině za druhého císařství, *Rougon-Macquartech*). Od společenského fatalismu vede pak další vývoj k fatalismu individuálnímu.

Když byla statistická šetření z oborů morální statistiky konána po delší dobu a když tento statistický materiál byl podroben zvláště *Lexisem* a *Bortkiewiczem* novým rozborům, došlo se k poznání, že pravidelnosti v časových řadách nejsou tak stálé, jak se z počátku jevily, nýbrž že i časové řady z oboru jevů morálních vykazují různé polyby a tendence jako jiné řady časové. Zjistilo se, že změny v hospodářských a sociálních poměrech mají vzápětí též změny v různých souborech t. zv. morální statistiky. Z toho bylo patrné, že nelze považovati pravidelnosti v souborech těchto jevů za neměnné, pevně do budoucnosti určené. Těmito novými poznatky není ovšem vyřešena základní sporná otázka, problém svobody vůle, neboť je možný též výklad mechanistický, že změny sociálních podmínek (na př. zvýšení životní úrovně dělnictva, zlepšení všeobecného vzdělání, rozšíření sociální péče, což jsou skutečnosti, které lze většinou sledovati též statisticky, nebo naopak zhoršení hospodářských poměrů v době hospodářských krisí) působí ve svém důsledku ony změny v poměrných četnostech jevů z oboru morální statistiky, které konstatujeme.

Pro problém individuální svobody lidské vůle (konkretisovaný na př. v otázku po odpovědnosti v trestním právu) nemůže statistická metoda poskytnouti ani kladného ani záporného argumentu, neboť problém ten je povahy filosofické.

**Literatura ke kapitole 12.** Vedle spisů v textu citovaných *Janko*, *Základy theorie statistické indukce*.

---

<sup>1)</sup> *Quételet* Sur l'homme et le développement de ses facultés, ou essai de physique sociale, Paříž 1835, *Adolf Wagner* Die Gesetzmässigkeit in den scheinbar willkürlichen menschlichen Handlungen, Hamburg 1864, *Buckle* The history of Civilization in England, Londýn 1858, sr. v české literatuře *Weyr* Problém svobody vůle a statistika, Sborník věd právních a státních, XI.



## GRAFICKÉ ZNÁZORNĚNÍ.

Statistické údaje, výsledky statistických pozorování, které podle své povahy jsou vyjádřeny v číslech, je možno, jako jiné číselné hodnoty, znázorniti též graficky t. j. pomocí geometrických obrazců, případně též pomocí jiných grafických pomůcek, jako map a kreseb.

Grafickým znázorněním lze dosáhnouti především větší názornosti a přehlednosti. Přímký a křivky nebo jiné geometrické obrazce představující rozložení souboru nebo časovou řadu jsou daleko názornější a jasnější než číselné údaje představující tytéž statistické soubory. Proto se používá grafického znázornění zejména tam, kde jde o popularisační účely statistické. Nevýhodou grafického znázornění je, že volba měřítka je ponechána na vůli statistikovi, a že tímto způsobem lze ovlivňovati dojem, kterým na pozorovatele působí.

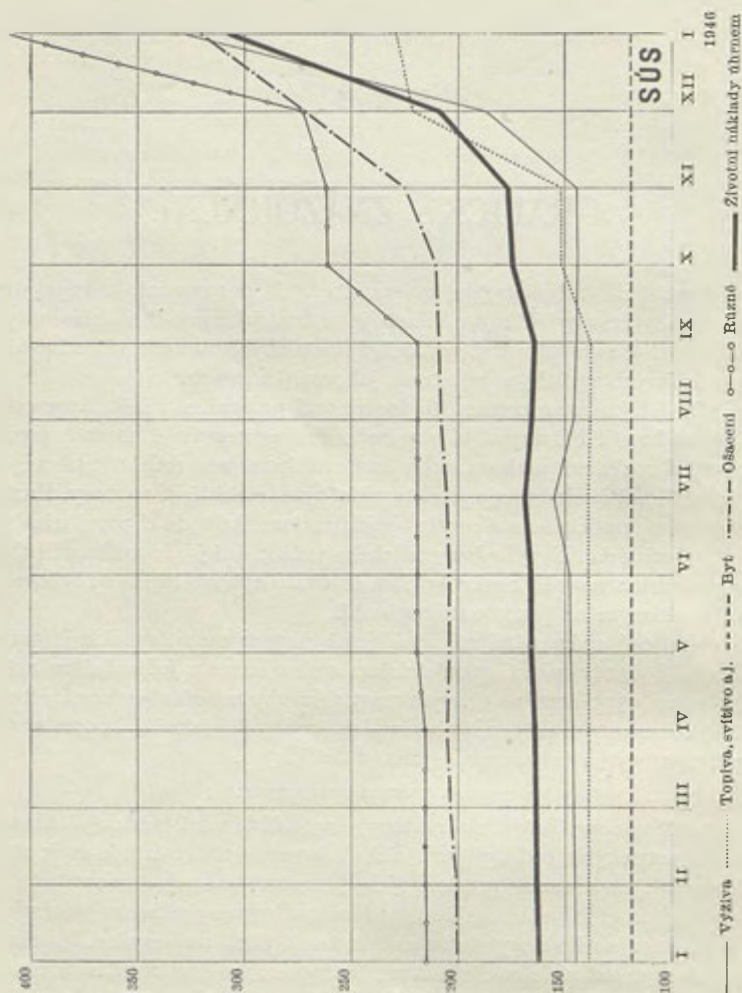
Grafické znázornění má však ještě jiný význam, umožňuje totiž některé hodnoty naléztí geometrickou cestou snáze a jednodušším způsobem než výpočtem, na př. vyrovnávati křivky rozložení četností, interpolovati statistické hodnoty v řadách časových, jakož i kontrolovati správnost výpočtů různých hodnot statistických.

Hlavní druhy grafického znázorňování jsou:

1. *Diagramy.* Statistické soubory se vyjadřují pomocí různých obrazců geometrických, bodů, přímek, křivek, čtverců, obdélníků, kružnic atd. Používá-li se obrazců prostorových (stereometrických), mluví se též o stereogramech. Nejčastěji jsou používány diagramy přímkové a křivkové, při kterých hodnoty statistické se nanášejí na dvě osy k sobě kolmé (v praxi se užívá obyčejně čtverečkovaného milimetrového papíru). Znázorňujeme-li graficky rozložení četností, nanášíme na osu horizontální

Příklad č. 35. Diagram (historiogram.)

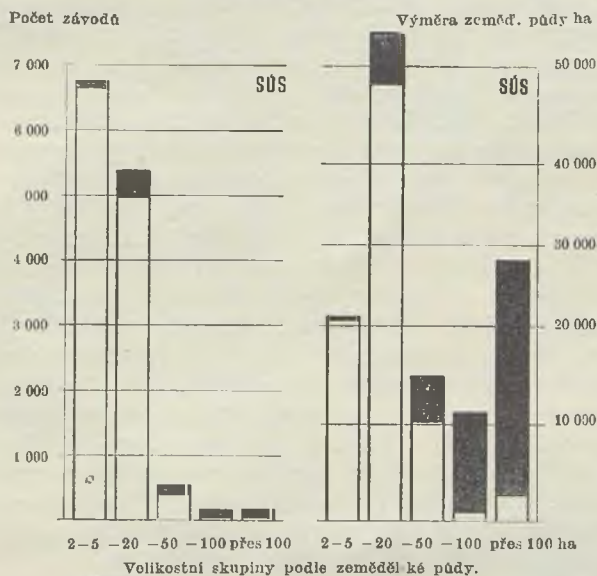
Vývoj indexů životních nákladů dělnické rodiny v Praze v r. 1945 a lednu 1946  
(základ III. 1939 = 100)



(osu  $x$ ) jednotky znaku (ev. třídy znaku), na pořadnice  $y$  (kolmice v těchto bodech vztyčené) příslušné četnosti znaku (tříd znaku). Spojením takto určených bodů přímkami nebo křivkami dostaneme diagram rozložení četností souboru (bývá nazýván též histogram). Podobně znázorňujeme řady časové nanášením časových jednotek na osu vodorovnou a příslušných hodnot na pořadnice (bývá nazýván též historiogram).

**Příklad č. 36. Diagram sloupkový.**

**Počet a výměra českých zemědělských závodů podle velikostních skupin v pohraničí r. 1939**



— zemědělské závody z pozemkové reformy

(Přehled Statistický zpravodaj roč. IX.)

Nespojujeme-li body takto získané na dvou osách kolmých čarami, nýbrž sestrojíme-li obdélníky spojením bodů, jež jsou vrcholy kolmic vztyčených ve středu příslušných intervalů, obdržíme t. zv. polygony četností.

Diagramy logaritmické jsou diagramy, sestrojené opět na dvou osách kolmých, na které se však nenanášejí hodnoty samé, nýbrž loga-



ritmy těchto hodnot. Logaritmický diagram podává jiný obraz statistické řady než diagram obyčejný a vyjadřuje poměr hodnot k sobě navzájem, zachycuje tedy vztahy hodnot, nevyjadřuje hodnoty samé. Diagramy logaritmické se hodí zvláště k znázornění řetězových indexních čísel.

Několika diagramy sestrojenými ve stejném měřítku na stejné síti lze sledovati i závislost mezi soubory a znaky. Nelze-li z technických důvodů použití měřítka stejného, užívá se též stupnice dvojitě, odehlné na levé a pravé straně diagramu.

Diagram sloupkový je sestrojen na stejném základě jako polygon četností, polygony se rozšiřují, aby bylo dosaženo větší názornosti, ve sloupky označené různými barvami nebo různě čárkované. Velikost hodnot je však vyjádřena toliko výškou sloupců.

Diagram osový nebo paprskový je diagram, kde hodnoty se vyznačují paprskovitě z jednoho bodu (středu kružnice), místo na dvě osy kolmé. Hodí se jen pro krátké statistické řady, na př. pro znázornění časové řady během dvanácti měsíců, hodnoty pro jednotlivé měsíce se nanášejí ze středu na dvanáct os.

Diagramy plošné (planimetrické) používají k znázornění statistických údajů obrazců plošných, nejčastěji čtverců nebo kružnic. Plošných diagramů používá se zejména tehdy, je-li třeba znázorniti hodnoty velmi veliké nebo navzájem velmi se lišící, takže bylo by obtížné vyjádřiti je měřítkem o jedné dimenzi. Plochy (čtverce, výseče kruhové) musí ovšem vyjadřovati přesně údaje a jejich poměr.

Diagramy prostorové (stereometrické) jsou znázorňovány obrazy prostorovými, krychlemi, koulemi, jehlany nebo válci. Jako obrazce o třech rozměrech mají výhodu, že se jimi mohou současně vyjádřiti tři rozměry, na př. korelační plocha v případě dílčí korelace mezi třemi znaky. Srovnání se zakládá na srovnání obsahů těles, sestrojování je dosti obtížné. Používá se jich též k účelům výstavním.

Podřadným druhem diagramů jsou diagramy obrázkové (piktogramy), kde se používá nikoli geometrických obrazců, nýbrž obrázků kreslených, na př. sudů, pytlů, figur lidských a pod. k vyjádření hodnot. Tyto diagramy slouží většinou pouze k účelům popularisačním a nebývají přesné.

2. *Kartogramy* (nazývané též statistické mapy) jsou zeměpisné mapy, na kterých se vyznačuje místní rozložení statistických hodnot pomocí různých barev nebo stínování. Takto je možno vyjádřiti různé stupně znaků kvantitativních nebo též různé obměny nebo různé stupně po-

Počet osob činných v průmyslu na 100 km<sup>2</sup> v zemi České a Moravskoslezské.

Příklad č. 37. Kartogram.



SÚS

-50
  51-150
  151-500
  501-2000
  2001-5000
  5001-15000
  extrémý

Oblastní jednotkou znárodnění jsou soudní okresy — Stav z počátku září 1945. (Pramen Statistický zpravodaj roč. IX.)



měrných četností znaků kvalitativních. V praxi je však nutno omeziti se na malý počet tříd nebo stupňů, jinak by se stal kartogram nepřehledným. Stupnice se vyjadřuje pomocí různých barev, nebo různou intenzitou téže barvy nebo různým stupněm stínování.

Kombinací zeměpisné mapy s diagramem vzniká t. zv. *kartodiagram*, v němž diagramy t. j. grafické znázornění statistických hodnot pomocí geometrických obrazců se lokalisují na mapě zeměpisné podle geografického rozložení hodnot.

**Literatura ke kapitole 13.** *Koller*, Graphische Tafeln zur Beurteilung statistischer Zahlen, Drážďany, 2. vyd. 1943.



## POMOCNÉ METHODY STATISTICKÉ.

Ve statistické praxi se často stává, že není možno zachytiti a statistickému šetření podrobiti celý soubor, tedy veškeré jednotky, z nichž se theoreticky soubor skládá. To se stává pro obtíže buď rázu technického nebo finančního. Není na př. možno zachytiti statisticky ceny všech obchodovaných statků při sestrojování cenových indexních čísel nebo celý soubor průmyslových výrobků při aplikaci statistické metody na zkoumání jejich jakosti.

V takových případech se používá ve statistické praxi method, které lze souborně označiti jako pomocné nebo náhradní metody statistické. Tyto metody jsou vesměs charakterisovány tím, že nezachycují celý soubor. Nejdůležitější tyto metody jsou:

### Částečné šetření methodou náhodného výběru.

Theoretické problémy částečného šetření methodou náhodného výběru byly naznačeny v kapitole 12. Ve statistické praxi je pak důležité, aby se výběr skutečně zakládal na náhodě, byl mechanický bez jakéhokoli subjektivního vlivu, abychom se vyvarovali systematických odchylek v tom nebo onom směru. Na př. koná-li se statistické šetření o souboru rodin, jejich počtu dětí, jejich spotřebě, vyberou se k šetření pouze rodiny, jejichž jména počínají určitým písmenem v abecedě, třeba A, H, P. Takový náhodný soubor je možno uspořádati též ze souborů, skládajících se z jevů okamžitých (událostí) tak, že se opět náhodně zvolí některé časové období k šetření, na př. určité dny (pátého, desátého, patnáctého) se vyberou k šetření frekvence na pouličních drahách.

Předpokladem pro správný úsudek, který se zakládá na úvaze pars pro toto, je stejnorodost souboru a dále výběr dosti veliký, aby se i ve vybraném vzorku mohl uplatnit zákon velkých čísel.

**Poznámka.** Určení rozsahu výběru při znaku kvalitativním a kvantitativním v. *Janko Jak vytváří statistika obrazy světa II*, str. 76 sl.

Z částečného souboru se usuzuje na celý soubor, jestliže naopak se z celého souboru usuzuje na jeho část, na př. z celoročního průměru cen zemáků se činí úsudek na průměrné lednové ceny zemáků, jak uvádí *Wagemann*, jde spíše o zvláštní případ odhadu. Zvláštní terminologie *Wagemannem* navržená pro tyto logické úsudky (jako substituce, inkluze a pod.) nenabyla rozšíření.

### Částečné šetření methodou výběru typických případů.

Částečné šetření statistického souboru lze provést též tím způsobem, že se záměrně vyberou případy charakteristické, typické pro celý soubor a na tomto výběru se vykoná statistické šetření. Výsledky takto získané šetřením výběru typických případů se pak považují za representační pro celý soubor. Odtud se nazývá tato metoda částečného šetření též methodou representační. Kdežto při methodě náhodného výběru rozhodovalo hledisko objektivní, mechanické, při této methodě rozhoduje hledisko subjektivní, záměrně se vybírá to, co se považuje za typické pro soubor. Avšak právě výběr případů skutečně typických, charakteristických pro soubor, působí v praxi velké obtíže a nezdaří-li se, není výsledek šetření směrodatný pro celek. Nejčastějším případem použití této metody je statistika domácích účtů či budgetů.

### Odhad.

Odhadu, t. j. přibližného ohodnocení, používá se dosti často ve statistice jevů společenských, když soubor se vymyká přesnému pozorování a sčítání. Při tom může býti odhad přibližným oceněním souboru, učiněným na základě věcné znalosti souboru, na př. odhad národního důchodu, který se opírá o určité částečné (i číselně vyjádřené) znalosti věcné, na př. statistické údaje o zdaněných důchodech, nebo odhad počtu obyvatelstva tam, kde není konáno sčítání lidu, na podkladě znalosti počtu obyvatel velkých měst a některých krajin, nebo odhad jednoho souboru na základě znalosti druhého, který máme statisticky vyšetřen.

Druhým případem odhadu je ten, který se zakládá na odhadech průměrů, které se pak kombinují se skutečnými statistickými údaji. Jako

nejčastější příklad tohoto druhu odhadu lze uvést t. zv. statistiku sklizní; ta se zakládá na odhadech průměrných hektarových výnosů sklizňových, které podávají zpravodajové (jmenovaní znalci) v každé obci; tyto odhadnuté výnosy se pak násobí zjištěnými plochami osetými příslušnou plodinou v každé obci. Výsledky této metody se pak obvykle označují jako statistika sklizně, ačkoli nejde o přímé statistické zachycení celého souboru, které jest z technických důvodů nemožné, nýbrž o uvedenou metodu odhadovou; proto též hodnota této statistiky, pokud jde o spolehlivost, nemůže se rovnati statistickým šetřením skutečným. Jiný příklad použití této metody jest odhad hodnoty zahraničního obchodu (tam kde není zavedena metoda přímého zjišťování této hodnoty, t. zv. metoda deklarační, jako jest tomu u nás), při kterém průměrné hodnoty jednotlivých druhů zboží získané odhadem se násobí množstvím dovezeného a vyvezeného zboží. zjištěným statistickým šetřením zahraničního obchodu.

#### Anketa.

Anketou (z francouzského enquête = vyšetřování) nazývá se šetření, které se snaží objasniti některou otázku dobrými zdáními (expertisami), vyžádanými od odborníků, znalců a zájemců o tuto otázku. Anketa má ovšem ráz subjektivní, poněvadž tlumočí osobní mínění a názory účastníků. Výsledky ankety neposkytují obvykle materiál, který by bylo vyjádřiti číselně ve formě ve statistice obvyklé. Anketa se však podobá statistickému šetření v tom, že se obrací na pokud možno velký okruh účastníků, aby se tak poznalo mínění převládající. Dobrá zdání účastníků mohou ovšem obsahovati statistický materiál, získaný jiným způsobem, na základě různých šetření, který pak pomáhá osvětliti problém, o který při anketě jde.

Ankety se pořádají nejčastěji pro řešení otázek hospodářských (na př. bytová anketa, měnová anketa), sociálních i kulturních.

#### Jiné pomocné metody statistické.

V nedostatku statistických údajů o nějakém souboru sahá se někdy ještě k jiným pomůckám, které mohou takový soubor osvětliti. Jsou to na př. údaje číselné, které se na soubor vztahují, ale nejsou výsledkem statistického šetření; tak kursovní listky bursovní jsou důležitou pomůckou pro statistiku cenovou, různé číselné údaje a záznamy ve starých právních listinách jsou pomůckou pro historická bádání statistická.

Podle své povahy jsou ceny zaznamenané na kursovních listkách burs



cenných papírů nebo burs se zbožím obyčejně nejčastější ceny, které se při bursovních obchodech vyskytly a byly zaznamenány. Rozdíl od statistického šetření jest při tom ten, že nejsou všechny případy obchodů systematicky statisticky vyšetřovány, nýbrž že se zaznamenává jen určitý zlomek, který postačí k určení kursu; na bursách s cennými papíry bývají pak udávány na kursovních listech poslední ceny, za kterých obchody byly sjednány. Tyto *nesystematické číselné záznamy*, jak se takové údaje označují, slouží za základ ke statistice cenové, ačkoli nejde, jak bylo vyřčeno, o pravé statistické šetření.

U souborů skládajících se z jevů okamžitých, které nelze stále běžné statisticky zaznamenávat, a které vykazují stálý pohyb v obou směrech, přírůstek i úbytek, používá se t. zv. *methody salda*, t. j. určuje se rozdíl mezi dvěma časovými obdobími u tohoto souboru; na př. saldo vkladů uložených a vybraných za určité časové období ve spořitelnách nezjišťuje sice, kolik bylo průběhem tohoto období peněz uloženo a vybráno, ale zjišťuje rozdílem úhrnu na začátku a na konci tohoto období výsledek tohoto pohybu. Podobně se zjišťuje výsledek migrace v období mezi dvěma sčítáními lidu saldem pohybu migrace v tomto období.

Konečně uvádí se jako zvláštní případ t. zv. *šetření tabelární*, při kterém se nedělají záznamy o jednotkách souboru, nýbrž přímo se tvoří skupiny a ty se zapisují s příslušnými údaji v tabulky již upravené. Tato metoda má ten nedostatek, že údaje nelze již později kontrolovati, pokud jde o jednotlivé případy, ze kterých se skupiny skládají, ani jinak zpracovati, než jak již předem byly seskupeny.

Velmi sporné je, zda lze k pomocným methodám statistickým zahrnouti též methodu monografickou, jak činí někteří starší autoři, jako *Mayr* a *Žižek*. Při této methodě se vybere ze souboru jediný typický případ, který se pak podrobně popíše a analyzuje. Z výsledků této studie se pak usuzuje na celek. Z toho vyplývá jistá podoba této metody s methodou částečného šetření výběrem typických případů. Methody této hojně se používá v sociologii (známé jsou studie *Le Playovy* Les ouvriers européens. Paříž 1855, ze které vzešel základ ke statistice domácích rozpočtů dále *L'organisation de la famille*, Paříž 1870).

O všech pomocných methodách statistických platí, že jsou methodami náhradními a že poskytují údaje méně spolehlivé nebo neúplné. Přes to, jak již bylo uvedeno, statistická praxe se nemůže bez nich obejít. Při dalším zpracování a hodnocení takto získaného materiálu třeba však vždy na to pamatovati a neklásti jej na roveň údajům získaným přímým statistickým šetřením.

## Seznam značek a symbolů v knize stále užívaných. \*)

- $p$  značí theoretickou pravděpodobnost, že nastane případ  $A$   
 $q$  značí theoretickou pravděpodobnost, že nastane případ  $B$ , protichůdný případu  $A$   
 $\frac{m}{n}$  značí poměr pozorovaných případů vykazujících hledaný znak k celému pozorovanému souboru (čili empirickou pravděpodobnost neboli poměrnou četnost)  
 $x$  značí horizontální osu v grafickém znázornění a potom i znak, který na tuto osu nanášíme (na př. v korelační tabulce)  
 $y$  značí vertikální osu v grafickém znázornění a potom i znak, který na tuto osu nanášíme  
 $\sigma_0$  značí theoretickou střední odchylku  
 $\sigma$  značí směrodatnou (skutečnou) odchylku  
 $e$  značí základ přirozených logaritmů = 2·718281...  
 $\pi$  značí Ludolfovo číslo = 3·141592...  
 $x_1, x_2, x_3$  značí jednotlivé hodnoty (členy) souboru  
 $r$  značí počet hodnot souboru  
 $A$  a  $\bar{x}$  značí aritmetický průměr  
 $\Sigma$  značí součet  
 $f_1, f_2, f_3$  značí četnost (frekvenci) hodnot  
 $A_0$  značí prozatímní aritmetický průměr  
 $G$  značí geometrický průměr  
 $H$  značí harmonický průměr  
 $A_Q$  značí kvadratický průměr  
 $Me$  a  $\tilde{x}$  značí medián  
 $I$  značí indexní číslo  
 $J$  značí interval (rozpětí) třídy  
 $Q_1, Q_3$  značí kvartil první a třetí  
 $M_0$  značí modus  
 $\theta$  značí průměrnou odchylku  
 $\xi$  značí odchylku jednotlivých členů souboru od aritmetického průměru souboru  
 $I_Q$  značí odchylku kvartilů  
 $V$  značí koeficient disperse (variability)  
 $Q$  značí divergenční koeficient Lexisův  
 $s$  značí průměr serie  
 $Q_{as}$  značí koeficient asociace (Yule-ův)  
 $r$  značí koeficient korelace (Pearsonův)  
 $\sigma_A$  značí theoretickou odchylku aritmetického průměru  
 $\sigma_d$  značí theoretickou odchylku rozdílu  
 $\sigma_r$  značí theoretickou odchylku Pearsonova koeficientu korelace  
 $\eta$  značí korelační poměr  
 $X^2$  značí test Pearsonův.

\*) Označení použita v některých zvláštních vzorech, která jsou na příslušném místě vysvětlena a více se neopakují, nejsou v tomto seznamu obsažena.

## OBSAH

Kapitola 1.	<i>Pojem statistiky</i>	9
Kapitola 2.	<i>Stručný přehled dějin statistiky</i>	12
Kapitola 3.	<i>Logický podklad statistiky</i>	17
Kapitola 4.	<i>Theoretické rozložení souboru a zákon velkého čísla</i>	22
Kapitola 5.	<i>Technika statistického šetření</i>	31
Kapitola 6.	<i>Statistické znaky a rozložení statistických souborů</i>	41
Kapitola 7.	<i>Statistické charakteristiky</i>	51
Kapitola 8.	<i>Čísla poměrná a čísla indexní</i>	62
Kapitola 9.	<i>Řady časové</i>	67
Kapitola 10.	<i>Statistické závislosti</i>	73
Kapitola 11.	<i>Závislost mezi znaky kvantitativními (korelace)</i>	84
Kapitola 12.	<i>Náhodné výběry</i>	99
Kapitola 13.	<i>Grafické znázornění</i>	107
Kapitola 14.	<i>Pomocné metody statistické</i>	113

---



Dr. Cyril Horáček:

**RUKOVĚŤ STATISTIKY**

Vydal v lednu 1947 spolek československých  
právníků „Všehrd“ Praha I., Pařížská 27.

Vytiskla tiskárna „Práce“ v Praze.

Náklad 2000 výtisků.

Cena 65,— Kčs.