Jak vytváří statistika obrazy světa a života. I. díl

PROF. DR. J. JANKO

Jak vytváří statistika obrazy světa a života

I. DÍL

CESTA K VĚDĚNÍ SV. 2

CESTA K VĚDĚNÍ

PROF. DR. JAROSLAV JANKO

JAK VYTVÁŘÍ STATISTIKA OBRAZY SVĚTA A ŽIVOTA

I. DÍL

S 18 obrazci

Vyšlo jako 22. svazek sbírky
CESTA K VĚDĚNÍ
vydávané Jednotou českých matematiků a fysiků v Praze za redakce
Dra R. BRDIČKY, Dra F. VYČICHLA a Dra L. ZACHOVALA

1942

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

Jako si dnes nedovedeme představit život bez motorisovaných prostředků dopravních na pevnině, na vodě i ve vzduchu, tak není již možno žíti plným životem občana na dosaženém celkovém průměru kulturní úrovně bez umění statistického myšlení, stejně jako by to nebylo možno bez schopností čtení a psaní. Je proto veřejným zájmem především dobrá organisace statistiky, která je učleněným přehledem národního života a jeho existenčních podmínek. K jejímu naplňování je třeba školených pěstitelů statistiky ve všech velmi četných a rozmanitých oborech činnosti a statisticky vzdělané veřejnosti, která by mohla sledovat výsledky jejich prací a užívat jich k všeobecnému dobru. Mohutný rozvoj statistických metod v tomto století prostřel teprve statistice u společného stolu s ostatními vědami, které ji dříve považovali za pomocnici ve své domácnosti, kde v tomto postavení pracovala několik tisíciletí, nazveme-li statistickými akcemi některé soupisy obyvatelstva a jeho majetku u jednotlivých kulturních národů čtyři tisíce let př. Kr. Její obor se ovšem za tak dlouhou dobu velmi změnil, takže z původní nauky o státě — disciplina politico-statistica — odkud dostala své jméno, se vyvinula věda, která se zabývá netoliko hospodářskými a sociálními poměry ve státě, nýbrž poskytuje velmi vyhledávaný nástroj vědám přírodním i technickým. Umožnila podrobné měření složitých jevů populačních, nové vědní obory jako biometrika nebo ekonometrika vděčí za svůj vznik statistice; pokrok ve výrobě zemědělské, hromadná výroba průmyslová a obchod se dnes neobejdou bez podnikové statistiky a bez statistické kontroly jakosti výroby. Celé úseky národního hospodářství, jako pojišťovnictví na statistice přímo spočívají. Vitální statistika proniká do mediciny a prohlížíme-li lékařskou literaturu, přesvědčíme se, že preventivní medicina

bude brzo zahrnovati „preventivní" statistiku a lékaři budou užívati therapie ve světle směrodatné odchylky a Pearsonova kriteria. Statistika proniká do psychoanalytických měření, jimž podléhají komplexy, záliby, energie, a dodala významu testu inteligence. To je jen několik pestrých ukázek, které mohou osvětlit význam Pearsonovy věty v Galtonově životopise, kde prohlásil, že je stejně důležito vyučovat logice statistiky jako analyse matematiky. Vniknutí do vědeckých základů statistiky vyžaduje určitých znalostí matematických a ovládání technického slovníku užívaného ve statistice čili nově vytvořené terminologie. Aby byl již co nejúplněji sňat lesk duchaplnosti s úsloví, že „statistikou lze všechno dokázat" nebo „statistika je přesný součet nesprávných čísel" je třeba popularisace elementární statistiky a proniknutí jejího do nejširších kruhů hlavně tím, že budou metody co nejvíce normalisovány a podány na zjednodušených typech. Ujal jsem se proto rád napsání tohoto svazku Cesty k vědění, který má pojednati o základních metodách statistiky, použitelných ve všech oborech, maje na zřeteli také potřebu převésti statistiku z ryzí odbornosti k účinné službě na velkých úkolech národního společenství.

Nejedná se v tomto svazku pouze o metodách, jimiž se vytvářejí statistické obrazy, nýbrž také o metodách tvořících nástroj, který zbystřuje statistický pohled na obrazy, z nichž teprve tak lze načerpati hlubšího poučení. Byla a je celá stupnice statistických hříchů. Proto je třeba, aby je bylo co nejvíce umožněno rozeznávat a poznávat, které vznikly z nevědomosti nebo i ze zlomyslnosti. Nesmíme se ovšem domnívati, že již užíváním matematiky nebo jednoduchých výpočtů je zabezpečena větší správnost a přesnost odvozených výsledků. Jak může býti nebezpečno statistickému postupu použití jen matematického myšlení, osvětluje se někdy tímto vtipem: Potřebuje-li 1 muž 120 dní, aby vystavěl domek, pak s ním musí býti hotovo 12 mužů za 10 dní a podle toho dále 120 mužů za 1 den, 960 mužů za hodinu, 57 600 mužů za minutu. To je tedy obrázek, před-

stavující, jak by mohl vésti automatický postup myšlení ad absurdum.

Čtenář, který si osvojí základní poznatky z tohoto svazku, bude moci se zdarem pokračovati ve studiu dalšího, obsahujícího teorii a praksi náhodného výběru kvantitativního znaku a representativní metody vůbec.

Odkazy na literaturu, vzadu uvedenou, jsou provedeny v textu čísly v lomených závorkách. Tabulky integrálu Laplaceova a exponentiely Poissonovy, potřebné k některým výpočtům, najde čtenář v knihách citovaných v seznamu literatury, zvláště v [6] a [8].

Děkuji p. doc. dr. Fr. Vyčichlovi, redaktoru této sbírky, za laskavé opatření obrázků a Jednotě českých matematiků a fysiků za úpravné vydání knihy.

V Praze v květnu 1942.

Jaroslav Janko.

ČÁST I.

(1,1) Hromadné pozorování je praktickou cestou k poznávání.

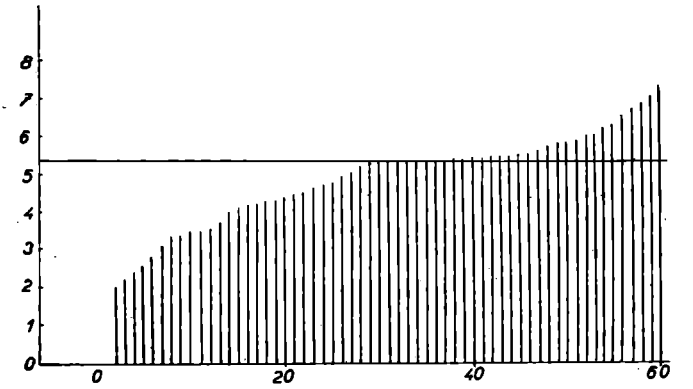
Chce-li majitel nákladního automobilu určité výrobní značky a typu znáti dobu života tohoto svého vozu, musí míti možnost změřit časový interval od okamžiku, kdy byl vůz dán jako hotový výrobek do provozu, do okamžiku skončení jeho poslední jízdy. Tento časový interval bude pro každý jednotlivý vůz téhož typu jiný, takže rozmanitost výsledných čísel nám nedá přesnou odpověď na další otázku, jaká je doba života vozu tohoto typu vůbec.

Podle získané zkušenosti přiřazujeme ke každé radioaktivní substanci určité číslo (konstantu) znamenající poměrnou část atomů, které se rozpadnou v nějakém daném čase. Tento rozpad se pozoruje pomocí fluoreskujícího stínítka. Vezme-li se na př. pro radium za jednotku času den, je zmíněné číslo přibližně jedna miliontina. Určité množství radioaktivní substance klesá tedy a zredukuje se na polovinu po době úměrné zmíněné konstantě, která je pro tu substanci charakteristická. Tato pravidelnost se projevuje vzhledem k tomu, že i nejmenší částečka substance obsahuje veliký počet atomů a lze ji potvrditi tím lépe, čím je pozorován vyšší počet rozpadů. Můžeme-li pozorovati přístrojem jen malý zlomek těchto rozpadů, pozorujeme přibližnou stálost. Nemůžeme však dáti přesnou odpověď na otázku, kolik atomů se rozpadne v příští minutě, neboť ten počet se mění; ale je tu jistá pravidelnost t. zv. statistická. Pro ni hledáme vyjádření.

Utrhneme-li nějaké množství vyspělých listů košatého stromu a změříme jejich délku, nebudou všechny stejné; mezi nejkratším a nejdelším bude třeba značný rozdíl. Opět bychom nemohli dáti určitou odpověď na otázku, jak dlouhý

list má dotyčný strom. Proto hledáme způsob, jakým bychom mohli dáti odpověď, která by jasně vyjádřila výsledky našich měření.

Seřadíme je tedy nejprve podle naměřené délky od nejmenšího do největšího a zobrazíme tuto posloupnost úsečkami v obr. 1 stejně od sebe vzdálenými. Shledáme, že konce



Obr. 1. Délka 60 listů s vyznačeným mediánem.

úseček dávají určitý průběh, zvláště kdyby byly spojeny, který svědčí tomu, že změny délek vytvářejí sled naznačující jistou pravidelnost. Chceme se přesvědčit zda tento obraz bude jiný, vezmeme-li v úvahu jiné množství listů téhož stromu. Provedeme tedy s druhým množstvím týž postup a shledáme, že průběh v této posloupnosti je zcela podobný; opět úsečky kolem prostřední, která má skoro touž délku jako v prvním případě vykazují od ní malé odchylky a směrem k oběma krajům je menší počet větších odchylek. Prostřední člen této posloupnosti se nazývá medián a lze jej při lichém počtu členů $2k+1$ určit jako $k+1$ člen od kraje, při sudém počtu $2k$ jsou dva prostřední členy (k-tý od každého

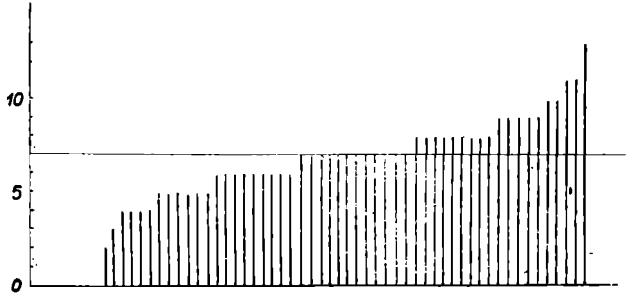
kraje) a hodnotu mediánu určíme jako jejich průměr. V našem zobrazení je jeho velikost vyznačena rovnoběžkou s osou základní. Tak se přesvědčíme, zvláště kdybychom tento postup opakovali vícekrát, že velikost mediánu a zobrazený tvar průběhu délek je pro měřený druh předmětů příznačný čili charakteristický. Ale velmi podobné tvary bychom dostali, kdybychom si znázornili výsledky měření proměnlivého znaku jiných předmětů, neboť jsme svůj příklad zvolili zcela nahodile. Jsme pak vedeni k domněnce, že za těmito pozorovanými jevy je nějaká všeobecná pravidelnost, která nám může pomoci k zjednodušení představ o velikosti hodnot znaku nějakého druhu předmětů.

Zvolený příklad dává tušiti jisté společné podmínky, za nichž vyrůstají listy téhož stromu a vytvářejí jejich délku, takže jsme tu ochotni očekávati nějakou zákonitost. Pokročíme proto dále ve své úvaze, budeme-li pozorovati nějaký jev, který patří mezi t. zv. náhodné, jako je házení kostkou nebo mincí. Hodíme do výše čtrnáct stejných mincí a po dopadu stanovíme počet rubů. Provedeme to třeba 201krát a výsledky, které jsme dostali jsou tyto:

7, 9, 7, 10, 6, 7, 5, 6, 7, 7, 9, 5, 7, 9, 8, 10, 7, 7, 6, 5,
8, 8, 8, 7, 8, 11, 8, 9, 5, 5, 7, 6, 4, 11, 6, 8, 10, 2, 5, 7,
7, 8, 3, 6, 8, 6, 7, 9, 8, 8, 6, 8, 8, 5, 7, 7, 4, 7, 9, 10,
5, 7, 8, 4, 7, 10, 11, 7, 5, 7, 6, 8, 8, 8, 9, 10, 7, 9, 7, 8,
8, 6, 7, 7, 7, 7, 5, 5, 7, 6, 7, 8, 9, 5, 7, 4, 5, 8, 7, 4,
4, 5, 5, 8, 7, 11, 7, 9, 5, 7, 8, 9, 8, 4, 10, 5, 5, 9, 4, 6,
7, 8, 7, 4, 9, 7, 13, 6, 4, 7, 6, 6, 9, 6, 7, 4, 6, 9, 6, 7,
6, 5, 6, 11, 6, 4, 5, 8, 7, 6, 10, 4, 9, 6, 8, 4, 8, 8, 7, 9,
9, 8, 10, 7, 7, 5, 6, 6, 7, 8, 6, 6, 8, 8, 11, 3, 4, 5, 7, 4,
6, 9, 4, 6, 6, 7, 8, 7, 4, 6, 5, 11, 6, 8, 11, 6, 3, 7, 6, 9, 7.

Tak jako jsme seřadili listy podle délky, tak seřadíme nyní pokusy se čtrnácti mincemi podle počtu rubů a výsledek si zase zobrazíme (obr. 2). Každá úsečka kromě krajních případů zastupuje přibližně čtyři výsledky. Dostá-

váme podobný tvar jako v dřívějším případě, takže přicházíme k myšlence, že by nám snad bylo usnadněno zkoumání takových případů, kdybychom odvodili teorii vysvětlující výsledky pokusů s jevy náhodnými. Tuto myšlenku budeme sledovati později.



Obr. 2. Počet rubů na 14 mincích.

(1,2) Hromadný jev. Nyní si především dobře všimneme, že k získání určitých poznatků nám nestačí pozorování, po případě měření jednoho předmětu (jedince) určitého druhu, nýbrž je třeba nahromaditi pozorování většího množství předmětů téhož druhu (prvků) čili nastoupiti cestu hromadného pozorování, což je vlastní metoda statistiky. Z jevu pozorovaného na jednotlivých předmětech téhož druhu se skládá jev hromadný, který vzniká působením určitého výseku všeobecných podmínek dění. Hromadný jev lze tedy pozorovati na souboru množství prvků odpovídajících určitému pojmu.

(1,3) Statistický soubor. Abychom sestrojili potřebný statistický soubor odpovídající určitému pojmu (člověk, dům, podnik, motorové vozidlo, strom, jabloňový list, úmrtí, sňatek), musíme definovat statistickou jednotku. Vymezíme tudíž nejprve věcně statistický soubor tím, že vytkneme podstatné znaky, které musí míti každý prvek, který má

býti zahrnut do souboru. Tak přejdeme od skutečných předmětů nebo jevů vnějšího světa k myšlenkovému předmětu (statistické jednotce) tím, že vytkneme znaky, jež považujeme s hlediska cíle šetření za podstatné; užijeme tedy logického postupu zvaného abstrakce.

(1,4) Statistická jednotka. Statistická jednotka vykazuje znaky a) shodné, obsažené v obsahu pojmu, které jsou společné všem prvkům, jež budou pojaty do souboru;

b) vyšetřované, které se u některých prvků vyskytují a u druhých nikoliv (znaky alternativní) nebo se u nich vyskytují v různém stupni (znaky kvalitativní nebo kvantitativní);

c) ostatní, které jsou jednak postižitelné, jednak nepostižitelné a mohou býti všem prvkům souboru společné, ale nemusí.

Definice statistické jednotky musí obsahovati znaky shodné, na jejichž základě je sestrojen zkoumaný statistický soubor, který je vzhledem k nim stejnorodý (homogenní) [1]. Není však stejnorodý vzhledem k vyšetřovaným znakům a k některým ostatním znakům. Tak na př. soubor lidí není stejnorodý vzhledem k znaku (alternativnímu) pohlaví nebo vzhledem k znaku (kvantitativnímu) věk. Můžeme však zkoumaný statistický soubor roztříditi podle některého vyšetřovaného znaku a tak z něho odvoditi nové částečné soubory, jejichž prvky mají kromě shodných znaků původního souboru ještě jeden nebo několik dalších shodných znaků (muži třicetiletí, ženy dvacetileté). Tyto soubory jsou stejnorodější než původní soubor, neboť množina shodných znaků je u nich větší. Jsou tedy různé stupně stejnorodosti čili formální homogenity statistického souboru, jejíž mez je dána především cílem dotyčného statistického šetření.

(1,5) Statistické číslo. Statistika se snaží udati, jako jeden z prvních výsledků statistického šetření, kolik prvků je zahrnuto definicí statistické jednotky čili jaký je rozsah souboru odpovídajícího určitému pojmu. Tak vzniká ze

statistických jednotek první statistické číslo. Tato statistická čísla osvětlují nejdůležitější skutečnosti lidského života ve společnosti a státě, jakož i poměry v přírodě; patří tudíž k základnímu stavu lidského vědění.

Pro hodnocení významu statistických čísel je důležito, abychom měli na paměti:

1. Prvky statistického souboru jsou vzájemně vázány pojmovým společenstvím, které je vyznačeno určitým stupněm stejnorodosti; homogenita je tedy proměnná.

2. Kromě pojmového společenství nemají prvky souboru vzájemně jiné vazby; jsou tedy na sobě nezávislé. Mezi jednotlivými prvky není strukturních vztahů, které však jsou mezi soubory (jednak mezi částečnými navzájem, jednak mezi nimi a původním).

3. Čím je stupeň homogenity vyšší, tím je zpravidla rozsah souboru menší.

(1,6) Statistika. Můžeme nyní také říci, co rozumíme statistikou, abychom vyznačili zorný úhel dalších výkladů.

Statistika je věda, jejímž předmětem je statistický soubor. Je to věda empirická, jejíž jednotnost je založena jednotným způsobem hromadného čili kolektivního pozorování vztahujícího se na množství předmětů nebo událostí.

Jako každá věda empirická má jednak úkol povahy materielní, který plní ve své části popisné a používá k tomu své techniky šetření a zpracování pozorovaných dat, jednak úkol povahy logické, který plní v části teoretické. Tato část musí odvozovati prvky pro své základy ze zkušenosti, aby logická konstrukce měla význam praktický. Spojení části abstraktní se skutečností tvoří věta o stálosti statistických četností, odvozená ze zkušenosti, jíž se budeme později blíže zabývati.

Můžeme rozeznávat dva obecné typy hromadného pozorování.

a) První typ je představován posloupností čísel $x_1, x_2, \ldots$ $\ldots, x_r$, která jsou výsledky pozorování resp. měření téhož

znaku na r předmětech či událostech téhož druhu; jedná-li se o znak alternativní, pak zjišťujeme u každého prvku toliko má-li tento znak nebo nemá a označujeme obyčejně jeho přítomnost číslicí 1 a nepřítomnost číslicí 0, takže výsledkem pozorování je sled jedniček a nul (na př. 010011101....).

b) Druhý typ je rovněž představen posloupností čísel $x_1, x_2, \ldots, x_r$, která však jsou výsledky r měření téhož znaku na jednom předmětu. Tyto číselné údaje shromažďujeme, chceme-li si opatřiti pro určitou ověřovací metodu odhad chyby měření.

Přesto, že metody zpracování výsledků obou typů hromadného pozorování jsou obdobné, zabývá se druhým typem zvláště teorie chyb [7]. Při našich úvahách budeme míti na mysli první typ hromadného pozorování.

(2,1) Technika statistického šetření a výsledek jeho v nashromážděných datech.

Číselné údaje získané statistickým šetřením čili hromadným pozorováním, nazýváme také statistická data nebo statistický materiál jakožto souhrn všech záznamů o prvcích zahrnutých do souboru; záznamy musejí býti metodicky bezvadné, aby bylo účelno vyvozovati z nich další úsudky pomocí statistické teorie, jejíž pevné základy tvoří metody, o nichž budeme dále jednati. Proto musí odpovídati logickým podkladům, vyloženým v předcházejícím oddílu, vyspělá technika statistického šetření, a to nejen rozsáhlého, jakým je sčítání lidu nebo sčítání závodů, statistika zahraničního obchodu, statistika mezd a pod., nýbrž i menšího rozsahu jako při studiu souboru nějakého počtu lebek, nebo nevelkých vzorků předmětů výroby průmyslové či zemědělské. Cesta získávání statistických čísel bývala prvním zdrojem „statistických hříchů“. Musí proto odborně a přesně připraviti souhrn všech definic a předpisů, jakožto rovinné zrcadlo, v němž se má hromadný jev číselně zobraziti, aby pokud možno nic nedeformovalo.

Každému statistickému šetření musí předcházeti jeho příprava. Má býti formulován účel šetření čili stanoven úkol nebo problém, který má býti dotyčným šetřením řešen nebo objasněn. (Na př. zjištění velikosti a složení obyvatelstva podle řady hledisk.)

(2,2) Plán šetření povahy logické. S hlediska účelu šetření musí býti vypracován plán šetření jednak pro získání (sbírání), jednak pro zpracování statistického materiálu. Plán šetření musí obsahovati:

1. Vymezení statistické jednotky a to věcné, prostorové a časové. Na př. má-li se zjistiti při sčítání lidu soubor přítomného obyvatelstva na určitém území, je to obyvatel na určitém územním prostoru v kritickém okamžiku (třeba o půlnoci na 1. prosince) přítomný. Znaky shodné mají býti udány tak, aby bylo možno v každém konkretním případě rozhodnouti o tom, patří-li do souboru či nikoliv. Časové vymezení je dáno určitým okamžikem při zkoumání t. zv. jevů trvalých, a určitým časovým intervalem u jevů okamžitých (porod, sňatek, úmrtí, α-částice vysílané radioaktivním zářením). Věcné vymezení se setkává s obtížemi v tom, že pojmy, jak je dává život, nejsou vždy jasné a jednoznačné (nezaměstnaný) a pojmy, které dává věda nebo právní řád, nejsou vždy všeobecně přijímány, ale bývají často sporné (reálná mzda). Proto musí statistik k určitému účelu sestrojiti často sám definici jednotky k určení pojmu. To je možné při provádění statistiky primérní, t. j. statistiky určené jen k poznání pozorovaných hromadných jevů. Musí tedy definovati pojmy podle možnosti tak jak to vyžadují příslušné vědní obory, technika, národní hospodářství, a pod.; kde však to vyžaduje účel šetření, může se odchýliti a vlastním zásahem některé skupiny zahrnouti či vyloučiti. Při zpracování musí ovšem dbáti toho, aby dal příslušné vysvětlení o pojmovém vymezení a tím také správný význam získaným číslům. Tak na př. při sčítání domů je nutno určiti, je-li jednotkou dům ve smyslu jednotky stavebně

technické nebo administrativní a v definici musí býti rozřešeno, jak se zařadí nouzové kolonie bez popisných čísel, weekendové chaty, obývané baráky a pod. Koná-li se šetření za účelem fiskálním, musí tvořit základ definice jednotky statistické, definice zákona o dani domovní. Je zřejmo, že v každém případě bude okruh zahrnutých případů, a tedy soubor, jiný. Tam, kde má býti statisticky zpracován materiál, který byl původně určen jinému účelu (na př. správnímu), vzniká statistika sekundérní. Zde na rozdíl od primérní statistiky je statistik vázán materiálem, jehož pojmy musí převzíti a nezná vlivu na jejich vytváření; musí je jen jasně uvésti, aby každý statistický spotřebitel tyto základní skutečnosti znal a musil jich dbáti při posuzování čísel a porovnávání s výsledky jiných statistik.

2. Stanovení vyšetřovaných znaků (na př. pohlaví, věk, povolání, národnost). Rozhodným je splnění účelu šetření, který někdy sdružuje řadu zájmů. Při statistice požárů na př. má věda hospodářská hlavně zájem na škodách vzniklých národnímu hospodářství, požární pojišťovny potřebují zjistit pro výměru pojistné prémie četnost požárů a řadu vlastností poškozených předmětů, aby mohly sestrojit nebezpečenské třídy.

3. Určení, které částečné soubory jest odvoditi z původního souboru. To vyžaduje stanovení podrobností o tom, podle kterých znaků má býti původní soubor roztříděn a v jakých kombinacích.

4. Rozhodnutí, která statistická čísla mají býti vypočtena (absolutní, relativní, různé charakteristiky, s nimiž se dále seznámíme) a do jakých tabulek budou seřazena a publikována.

(2,3) Plán organisačně technický pro sbírání a zpracování materiálu. Tato rozhodnutí povahy logické jsou pak doprovázena opatřeními organisačně technickými pro sbírání a zpracování statistického materiálu. Podle statistické jednotky a vyšetřovaných znaků jakož i hledisek celého zpracování se vypracuje dotazník nebo sčítací arch, v němž

jsou položeny všechny potřebné otázky co nejjasněji se zřetelem k osobám, které jej budou vyplňovat, aby odpovědi byly snadné a správné. Obyčejně se připojuje také návod k vyplňování formulářů čili instrukce. Šetření se provádí buď individuálními lístky, t. j. pro každý prvek souboru, nebo hromadnými, tedy sběrnými listinami pro celé skupiny (domácnost, bytová strana), které nabývají značného rozšíření vzhledem k zpracování pomocí elektrických strojů, poněvadž usnadňují běžné vyznačování znaků podle předem vypracovaných klasifikačních schemat a dírkování, kterým se přenášejí znaky na tuhé štítky vhodné k třídění. Třídění se může prováděti při malém rozsahu souboru pomocí čárkování, lepení známek z útržkového bloku nebo vkládáním sčítacích lístků čili štítků, při velkém rozsahu pomocí strojů [3].

Musí býti stanovena dále úplná organisace pro způsob, jakým se provede sbírání a zpracování. Rozdělí se vhodným místům a orgánům úkoly rozeslání dotazníků, vyplnění, revise, sbírání a kontroly buď s hlediska statistického zpracování centrálního, při němž se veškerý materiál shromáždí v jednom místě nebo decentralisovaného v několika místech. Moderní potřeby vyžadují při rozsáhlejších šetřeních množství znakových kombinací. To vyžaduje složitého zpracování pomocí vyspělé techniky, které může býti provedeno jen dobře vyzbrojenými statistickými centrálami, jež musejí všechen původní materiál do svých rukou převzíti a přesvědčiti se o jeho úplnosti. Nespadá do rámce našich úvah o obecných metodách výklad podrobností technického procesu, který musí býti vždy přizpůsobován konkretnímu šetření a studován z učebnic k tomu cíli sepsaných jako [2] a [3]; do přípravy plánu určitého šetření patří také studium metody a techniky užité při takovém šetření snad již dříve nebo jinde a posouditi jak dobře vedla k zamýšlenému cíli.

(2,4) Plán publikační. Vyložili jsme si metodický postup, jímž vznikají statistická čísla, která se pak jako výsledky statistického šetření po přezkoumání číselné správ-

nosti a vnitřní shody sestavují do číselných přehledů čili tabulek nejprve soustředných. Z nich se podle publikačního plánu odvozují hlavní tabulky podrobné, pak přehledné, srovnávací a jiné. Pro formální úpravu jednotlivých tabulek, jež jsou vyznačeny nápisem, hlavičkou seřazující logicky sloupce a legendou popisující obsah řádků, platí důležité zásady, jejichž splnění vyžaduje poslání tabelárních přehledů [2]. Výsledky rozsáhlých šetření statistických úřadů jsou publikovány v publikacích jednak periodických, jednak v obsáhlejších dílech pramenných [3]. Vyskytuje se také nový systém „mikroskopických archivů". Podle něho se obsáhlé tabulky, jichž je však třeba k nějakému účelu často užívati, fotografují na velmi malou plošku, takže se sto čtverečních jednotek tabulky psané nebo tištěné objeví na třech čtverečních jednotkách filmu. Tabulky tohoto mikroskopického archivu se promítají, aby je bylo možno čísti. Takový postup urychlí některé práce užívající tohoto materiálu a chrání jej před opotřebováním.

Ke konci tohoto odstavce třeba ještě zvlášť zdůrazniti, že jakmile vyjádříme rozsah souboru určitým číslem, které se vztahuje na prvky s jistou množinou společných znaků, zmocní se potom tohoto čísla matematika. Při jejích operacích zmizel empiricko-statistický význam tohoto čísla. Kdyby se jednalo na př. o pojem „hodiny" definovaný jako „zařízení k měření času", jehož obsah je tedy dán dvěma znaky, zahrnoval by hodiny sluneční, přesypací, kyvadlové věžní, pokojové, kapesní, náramkové ve všech rozmanitých druzích atd. Tomuto pojmu odpovídá jistý statistický soubor, jehož rozsah je dán statistickým číslem, složeným ze statistických jednotek, které mají dva znaky společné, v ostatních pak ponechávají velikou rozmanitost; a tyto statistické jednotky se tu sčítají. Zpracování matematické přihlíží k takovému číslu jako k ryzímu číslu, kde mezi jednotkami není vůbec rozdílu. Proto musí statistik při rozboru výsledků na konec viděti vždy za číslem jasně jeho podklad empiricko-statistický.

ČÁST II.

(3,1) Metody k zhuštění informace vyjádřené posloupností původních dat. (Seřazení a úprava materiálu. Variační obor. Kvartily.)

Sebráním materiálu jsou prvky zkoumaného empirického souboru zastoupeny dotazníky nebo sčítacími lístky, v nichž jsou jejich vyšetřované znaky zapsány a na ně se vztahuje další zpracování. Údaj, který popisuje určitý znak se nazývá statistická proměnná (x). Je-li znak určen jednou proměnnou, nazývá se jednorozměrným, jinak vícerozměrným. Zjišťujeme-li údaje o vyšetřovaných pracích k určitému okamžiku čili studujeme stav souboru v něm, přihlížíme tedy k statické stránce problémů. Určujeme-li časové změny znaků a sledujeme tak změnu příslušnosti prvků do souboru v čase čili kinematiku souboru, přihlížíme k dynamické stránce problémů. V dalším se budeme zabývati jen statickou stránkou.

Pozorováním určitého kvantitativního znaku jsme na př. dostali v jednotkách míry (nebo váhy atd.) konkretní posloupnost hodnot odpovídajících rozsahu $r = 270$ prvků souboru

101	140	78	63	138	110	90	135	58	89
110	102	80	102	99	98	96	110	106	70
103	122	92	107	111	118	106	125	108	103
87	95	140	74	124	80	80	82	88	114
101	118	86	101	84	57	107	107	70	100
88	82	101	86	80	117	97	97	107	115
109	102	83	103	115	84	89	110	92	74
112	99	110	73	133	106	108	97	83	151
83	82	83	105	27	94	66	103	110	104
138	129	123	119	115	98	87	97	132	83
86	86	82	118	100	134	99	75	81	109

118	74	107	87	46	117	80	88	87	92
88	102	69	99	83	67	110	99	91	85
71	92	103	91	98	131	102	110	108	120
80	139	102	76	118	89	84	86	89	92
111	83	124	78	161	148	104	96	130	86
85	108	80	104	65	104	87	108	102	78
115	119	118	79	92	110	91	72	95	114
123	107	96	97	100	91	163	86	86	89
85	105	117	106	94	87	94	123	92	124
73	115	90	103	138	95	138	106	88	107
85	98	94	101	108	117	119	95	97	109
105	136	78	109	86	82	112	127	89	133
97	81	101	76	126	96	98	90	114	73
110	94	88	118	113	100	91	111	90	100
89	110	100	82	79	108	136	98	126	113
116	101	71	70	201	124	89	115	93	86

Na těchto datech uvedených ve formě neuspořádané prvotní tabulky ukážeme metody seřazení a soustředění do menšího počtu čísel. Prvním krokem v úpravě množiny pozorovaných hodnot jedné proměnné x je seřazení jejich podle velikosti, a to v pořadí hodnot neklesajících $x_1 \leqq x_2 \leqq \ldots \leqq x_r$. Při tom tedy najdeme nejmenší pozorovanou hodnotu proměnné $x_1 = 27$ a největší $x_r = 201$, jejichž rozdíl $x_r - x_1 = 174$ se nazývá **variační obor** nebo **variační rozpětí**, a snadno najdeme medián $\tilde{x} = 99$, před nímž je v seřazené posloupnosti r' členů a za ním r'' členů, přičemž $r' = r''$. Vyjmeme-li takový člen před nímž je r' a za ním r'' členů tak, že $3r' = r''$ dostáváme dolní **kvartil** $\tilde{x}_1 = 86$ před nímž je tudíž čtvrtina členů posloupnosti a za ním tři čtvrtiny.

Je-li $r' = 3r''$, nazýváme příslušný vyňatý člen horní **kvartil** $\tilde{x}_2 = 110$. Tak rozdělují oba kvartily a medián pozorované hodnoty na čtyři skupiny o stejném počtu prvků.

Seřazení hodnot x:

27	78	83	87	92	97	102	107	110	118	131
46	78	83	88	92	98	102	107	110	118	132
57	78	83	88	92	98	102	107	110*	118	133
58	79	84	88	92	98	102	107	110	118	133
63	79	84	88	92	98	102	107	111	118	134
65	80	84	88	93	98	102	107	111	118	135
66	80	85	88	94	98	102	107	111	118	136
67	80	85	89	94	99	103	108	112	119	136
69	80	85	89	94	99	103	108	112	119	138
70	80	85	89	94	99*	103	108	113	119	138
70	80	86	89	94	99*	103	108	113	120	138
70	80	86	89	95	99	103	108	114	122	138
71	81	86	89	95	100	103	108	114	123	139
71	81	86	89	95	100	104	108	114	123	140
72	82	86	89	95	100	104	109	115	123	140
73	82	86	90	96	100	104	109	115	124	148
73	82	86	90	96	100	104	109	115	124	151
73	82	86*	90	96	100	105	109	115	124	161
74	82	86	90	96	101	105	110	115	124	163
74	82	86	91	97	101	105	110	115	125	201
74	82	87	91	97	101	106	110	116	126	
75	83	87	91	97	101	106	110	117	126	
76	83	87	91	97	101	106	110	117	127	
76	83	87	91	97	101	106	110	117	129	
78	83	87	92	97	101	106	110	117	130	

(3,2) Momentové charakteristiky (obecné, kolem aritmetického průměru, momenty směrodatné proměnné). Takové uspořádání hodnot má někdy svůj význam v počátečním stadiu rozboru. Nelze však ani při poměrně nevelkém rozsahu souboru, jakým je náš příklad, zachytiti v mysli takové množství čísel v celku, proto je třeba zhušťování. K němu spějeme dvojí cestou. První cesta spočívá v tom,

že si definujeme určité konstanty, které charakterisují takové posloupnosti. Snažíme se, aby byly definovány jednoduchým způsobem, aby byly snadno počitatelné a zahrnovaly všechny údaje.

Nejjednodušší charakteristikou, splňující tyto podmínky, je aritmetický průměr proměnné x, který se rovná součtu všech hodnot proměnné, dělenému jejich počtem. Označíme-li $\bar{x}$ aritmetický průměr nebo krátce průměr hodnot $x_1, x_2, \ldots, x_r$ pak tedy platí rovnice

$$\bar{x} = \frac{1}{r}(x_1 + x_2 + \ldots + x_r) = \frac{1}{r}\sum_{i=1}^{r} x_i. \qquad (1)$$

Aritmetický průměr se nazývá také prvním momentem. Zcela obdobně se pak definují další momenty, takže k-tý moment $\mu'_{x,k}$, který se také nazývá momentem k-tého řádu, jest průměrem k-tých mocnin hodnot proměnné, je tedy vyjádřen rovnicí

$$\mu'_{x,k} = \frac{1}{r}(x_1^k + x_2^k + \ldots + x_r^k) = \frac{1}{r}\sum_{i=1}^{r} x_i^k \qquad (2)$$

aritmetický průměr ovšem plyne z této rovnice pro $k = 1$, takže $\bar{x} = \mu'_{x,1}$ a další momentové charakteristiky dostáváme, klademe-li $k = 2, 3, 4, \ldots$

Vedle těchto obecných momentů mají ve statistice zvláštní význam momenty kolem aritmetického průměru.

Označíme-li odchylku jednotlivých hodnot proměnné od aritmetického průměru $\xi_i = x_i - \bar{x}$, potom momenty kolem aritmetického průměru $\mu_{x,k}$ definujeme

$$\mu_{x,k} = \frac{1}{r}\sum_{i=1}^{r} \xi_i^k. \qquad (3)$$

První moment kolem aritmetického průměru je roven nule, neboť pro $k = 1$

$$\mu_{x,1} = \frac{1}{r}\sum_{i=1}^{r} \xi_i = \frac{1}{r}\sum_{i=1}^{r} (x_i - \bar{x}) = \\ = \frac{1}{r}\sum_{i=1}^{r} x_i - \frac{r\bar{x}}{r} = \bar{x} - \bar{x} = 0. \tag{4}$$

Pro výpočet dalších momentů kolem aritmetického průměru pro $k = 2, 3, 4, \ldots$ mají význam vztahy, které platí mezi nimi a momenty obecnými. Snadno je odvodíme takto

$$\begin{aligned} \xi_1^2 &= x_1^2 - 2x_1\bar{x} + \bar{x}^2 \\ \xi_2^2 &= x_2^2 - 2x_2\bar{x} + \bar{x}^2 \\ &\ldots\ldots\ldots\ldots\ldots \\ \xi_r^2 &= x_r^2 - 2x_r\bar{x} + \bar{x}^2 \end{aligned}$$

$$\sum_{i=1}^{r} \xi_i^2 = \sum_{i=1}^{r} x_i^2 - 2\bar{x}\sum_{i=1}^{r} x_i + r\bar{x}^2$$

Odtud plyne dělením r

$$\frac{1}{r}\sum_{i=1}^{r} \xi_i^2 = \frac{1}{r}\sum_{i=1}^{r} x_i^2 - 2\bar{x} \cdot \bar{x} + \bar{x}^2$$

čili

$$\mu_{x,2} = \mu'_{x,2} - \bar{x}^2. \tag{5}$$

Obdobně

$$\begin{aligned} \xi_1^3 &= x_1^3 - 3x_1^2\bar{x} + 3x_1\bar{x}^2 - \bar{x}^3 \\ \xi_2^3 &= x_2^3 - 3x_2^2\bar{x} + 3x_2\bar{x}^2 - \bar{x}^3 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \xi_r^3 &= x_r^3 - 3x_r^2\bar{x} + 3x_r\bar{x}^2 - \bar{x}^3 \end{aligned}$$

$$\sum_{i=1}^{r} \xi_i^3 = \sum_{i=1}^{r} x_1^3 - 3\bar{x}\sum_{i=1}^{r} x_i^2 + 3\bar{x}^2\sum_{i=1}^{r} x_i - r\bar{x}^3$$

takže

$$\frac{1}{r}\sum_{i=1}^{r} \xi_i^3 = \frac{1}{r}\sum_{i=1}^{r} x_i^3 - 3\bar{x}\frac{1}{r}\sum_{i=1}^{r} x_i^2 + 3\bar{x}^2 \cdot \bar{x} - \bar{x}^3$$

a tedy

$$\mu_{x,3} = \mu'_{x,3} - 3\bar{x}\mu'_{x,2} + 2\bar{x}^3. \tag{6}$$

Stejně se odvodí

$$\mu_{x,4} = \mu'_{x,4} - 4\bar{x}\mu'_{x,3} + 6\bar{x}^2\mu'_{x,2} - 3\bar{x}^4. \qquad (7)$$

Dalších momentů se užívá velmi zřídka a obecný vztah se snadno najde rozvedením $(x_i - \bar{x})^k$ podle binomické věty [11, 12].

Druhého momentu kolem průměru si zvláště povšimneme, neboť spočívá na součtu čtverců odchylek proměnné od průměru a proto charakterisuje rozptyl pozorovaných hodnot proměnné. Obyčejně se užívá k měření rozptylu čili variability jeho druhé odmocniny, která se nazývá směrodatná odchylka

$$\sigma_x = \sqrt{\mu_{x,2}}, \quad = \sqrt{\frac{1}{r}\sum \xi^2} \qquad (8)$$

neboť pak je míra téhož rozměru jako pozorovaný znak. Při výpočtu vychází tedy v těchže jednotkách, v nichž jsou napozorované hodnoty proměnné a její čtverec $\sigma_x^2 = \mu_{x,2}$ nazýváme rozptyl.

Zavádíme ještě pojem „směrodatná proměnná" t_i tak, že měříme odchylky proměnné od průměru směrodatnou odchylkou čili vyjádříme je v jednotce „směrodatná odchylka", potom

$$t_i = \frac{x_i - \bar{x}}{\sigma_x} = \frac{\xi_i}{\sigma_x}, \qquad (9)$$

čímž dostáváme čísla bez rozměru, můžeme říci abstraktní; poskytují však výhody při mnohých matematických operacích a usnadňují některá srovnávání.

Významné jsou některé vlastnosti momentů směrodatné proměnné

$$\mu_{t,1} = \bar{t} = \frac{1}{r}\sum_{i=1}^{r}\frac{x_i - \bar{x}}{\sigma_x} = \frac{1}{\sigma_x}\frac{1}{r}\sum_{i=1}^{r}\xi_i = 0,$$

$$\mu_{t,2} = \frac{1}{r}\sum_{i=1}^{r}\frac{(x_i - \bar{x})^2}{\sigma_x^2} = \frac{1}{\sigma_x^2}\frac{1}{r}\sum_{i=1}^{r}\xi_i^2 = \frac{\mu_{x,2}}{\sigma_x^2} = 1,$$

$$\mu_{t,3} = \frac{1}{r}\sum_{i=1}^{r}\frac{(x_i - \bar{x})^3}{\sigma_x^3} = \frac{1}{\sigma_x^3}\frac{1}{r}\sum_{i=1}^{r}\xi_i^3 = \frac{\mu_{x,3}}{\sigma_x^3}.$$

Ačkoliv tedy $\mu_{t,1} = 0$, $\mu_{t,2} = 1$, hodnota třetího momentu směrodatné proměnné $\mu_{t,3}$ závisí na hodnotách proměnné. Je známa pod názvem šikmost nebo kosost a označuje se symbolem $\alpha_{x,3}$, takže

$$\alpha_{x,3} = \frac{\mu_{x,3}}{\sigma_x^3} = \frac{\mu_{x,3}}{\mu_{x,2}\sigma_x}. \tag{10}$$

Postoupíme-li dále ke čtvrtému momentu, vidíme, že

$$\mu_{t,4} = \frac{\mu_{x,4}}{\sigma_x^4} \tag{11}$$

a označujeme jej pak obyčejně $\alpha_{x,4}$. Je výrazem špičatosti nebo plochosti; užívá se ho k tomu účelu ve tvaru $\alpha_{x,4} - 3$ a nazývá se koeficientem špičatosti nebo excesem.

Seznámili jsme se tedy se základními momentovými charakteristikami souboru, jimiž jsou:

1. rozsah souboru r,
2. aritmetický průměr hodnot proměnné $\bar{x}$,
3. směrodatná odchylka σ_x,
4. šikmost (kosost) $\alpha_{x,3}$,
5. špičatost (exces) $\alpha_{x,4} - 3$.

Jejich praktický výpočet vyžaduje zjednodušení výpočtu momentů. K tomu lze užíti základního teorému o momentech, který zní: momenty kolem aritmetického průměru se nemění, zvětší-li se nebo zmenší-li se všechny hodnoty proměnné o stejnou konstantu.

Důkaz provedeme snadno, odečteme-li na př. ode všech hodnot proměnné x hodnotu x_0, které se někdy říká předběžný průměr. Potom je $x_i - x_0 = \eta_i$ nová proměnná, pro niž jsou charakteristiky definovány stejně jako pro x, čili průměr

$$\bar{\eta} = \frac{1}{r}\sum_{i=1}^{r}\eta_i,$$

a k-tý moment kolem průměru

$$\mu_{\eta,k} = \frac{1}{r}\sum_{i=1}^{r}(\eta_i - \bar{\eta})^k.$$

Především dostáváme

$$\sum_{i=1}^{r} x_i = rx_0 + \sum_{i=1}^{r}\eta_i \text{ čili } \bar{x} = x_0 + \bar{\eta}.$$

Můžeme tedy psáti

$$\eta_i - \bar{\eta} = (x_i - x_0) - (\bar{x} - x_0) = x_i - \bar{x} = \xi_i$$

a tudíž moment

$$\mu_{\eta,k} = \frac{1}{r}\sum_{i=1}^{r}(\eta_i - \bar{\eta})^k = \frac{1}{r}\sum_{i=1}^{r}\xi_i^k = \mu_{x,k}, \qquad (12)$$

což je důkazem, že se nezměnil.

(3,3) Tabelární podávání výsledků. Rozdělení četností. Druhá cesta ke zhuštění pozorovaných dat vede přes tabulku rozdělení četností.

Statistický soubor převedeme do tabulky rozdělení četností, která má dva sloupce. V prvním jsou seřazeny podle velikosti jen různé hodnoty proměnné, které byly pozorovány. V druhém sloupci je uveden počet prvků, na nichž byla každá z těchto hodnot při šetření zjištěna. Tomuto počtu prvků s hodnotou proměnné x_i říkáme četnost n_i.*) I tato tabulka rozdělení četností je ještě málo přehledná. Proto postupujeme dále ke skupinovému rozdělení četností.

(3,4) Skupinové rozdělení četností. Tento typ tabulky vzniká tak, že několik hodnot proměnné se sdruží k utvoření jednoho intervalu a četnost se uvádí jedna pro celý interval, zvaný třída. Tak máme v prvním sloupci třídní intervaly a v druhém sloupci celkovou četnost všech hodnot proměnné,

*) Čtenář si laskavě sám napíše tuto tabulku z čísel našeho příkladu na str. 19.

které spadají do intervalu. Tím že tato tabulka již nepodává četnost každé původní pozorované hodnoty, nepředstavuje je již přesně a něco z původní informace se ztrácí v zájmu přehlednějšího obrazu o všeobecném tvaru rozdělení četností.

V našem číselném příkladu sdružíme třeba hodnoty 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, do jednoho intervalu, který bude zastupován zpravidla jejich průměrem, zde 90, a počet prvků spadajících do této třídy bude 80, což je tedy třídní četnost n_i. Tento interval má délku 15 jednotek, v nichž byl znak měřen. Postupujeme pak podél stupnice měření a zachovávajíce stálou délku intervalu, rozdělíme celé variační rozpětí na stejné intervaly a celý soubor na třídy. Při této konstrukci jak patrno, vznikají dvě důležité otázky. Jednak třeba stanoviti délku čili velikost intervalu s čímž souvisí jejich počet, jednak vymeziti hranice intervalů. S tím souvisí také otázka počátku prvního intervalu. Řešení těchto otázek se poněkud liší dle toho, jedná-li se o znak rozpojitý, kdy proměnná nabývá jen isolovaných hodnot, nebo o znak spojitý, kdy probíhá všechna reálná čísla určitého intervalu. Nejprve přihlédneme k prvnímu případu.

(3,5) Délka a hranice třídního intervalu. Není-li specielních potřeb daných přímo účelem šetření, pak určují volbu velikosti třídního intervalu dvě všeobecné podmínky:

1. hodnoty proměnné, zařazené do třídního intervalu lze pokládati s hlediska cíle šetření za zastupitelné průměrnou hodnotou třídního intervalu, která je zpravidla totožná s prostřední hodnotou. (Chceme-li sestaviti na př. tabulku úmrtnosti nějakého souboru osob podle věkových skupin, rozhodneme se pro interval 3- nebo 5-letý, podle toho, stačí-li při použití této tabulky zastoupení úmrtnosti krajních věků intervalu pětiletého či tříletého úmrtností věku prostředního.)

2. Při zachování první podmínky má býti délka intervalu co největší. V praksi bývají tyto podmínky splněny nejčastěji volbou intervalu takové délky, že se soubory podle velikosti rozsahu rozpadnou do 10 až 20 tříd. Vycházíme-li z této zkušenosti, stanovíme přibližně délku intervalu, dělíme-li variační rozpětí $(x_r - x_1)$ počtem tříd, který je zvolen tak, aby při malém rozsahu souboru byly třídy obsazeny, čili měly dostatečnou četnost a při velkém rozsahu, aby bylo rozdělení četností přehledné.

Také lze určovati velikost třídního intervalu se zřetelem ke směrodatné odchylce, kterou k tomu účelu zhruba odhadneme z předpokladu, že variační rozpětí se přibližně rovná šestinásobku směrodatné odchylky (pravidlo šesti sigma [5]). Velikost třídního intervalu h potom určíme tak, aby splňovala nerovnosti $2h < (x_r - x_1) : 6 < 4h$.

Tato libovůle v určování velikosti třídního intervalu a tím počtu tříd je ovšem pro matematickou statistiku velmi nepříjemná. Jsou proto odvozeny také způsoby určování, spočívající na porovnání s binomickým rozdělením četností [str. 66. rovnice (37)].

Druhým úkolem je stanovení hranic intervalů. Při znaku rozpojitém se snažíme stanovit dolní a horní hranici intervalu tak, aby bylo o každé hodnotě proměnné jasno, do kterého intervalu patří. Obyčejně se stanoví v polovině mezi jednotkami posledního místa, v němž byly hodnoty proměnné uvedeny. V našem příkladu se jedná o čísla celá, takže interval svrchu zmíněný vyznačíme hranicemi 82,5 až 97,5. Střed intervalu, který zastupuje všechny hodnoty do něho spadající, zůstává tak číslo celé.

Zpravidla zavádíme všechny intervaly téže délky, ač materiál si někdy vynutí výjimky, zvláště tehdy, kdy by byl rozsah stupnice příliš veliký a některé obory na př. vysokých hodnot velmi řídce obsazeny. Také někdy první interval bývá dolu neohraničený, takže jsou do něho za-

řazeny všechny hodnoty až do stanovené horní hranice jeho; podobně je někdy poslední interval neohraničený nahoru. Poloha intervalu stanovená dolní hranicí bývá dosti libovolná a její volba nemá velkého vlivu na hodnoty charakteristik; někdy však vyplývá z povahy materiálu.

Kupí-li se na př. hodnoty pozorované nápadně kolem určitých čísel na př. 5 nebo 10, snažíme se, aby tato čísla padla do středu intervalu, který jej zastupuje. Hodnoty, které zastupují intervaly, budeme opět značit x_i, $i = 1, 2 \ldots$, takže hranice intervalů délky h budou

$$x_i - \tfrac{1}{2}h, \; x_i + \tfrac{1}{2}h.$$

V případě spojitého znaku nemůžeme uvésti všechny jeho hodnoty. Rozdělíme zase jeho celé variační rozpětí na l stejných částí a dostaneme intervaly $(x_i - \frac{1}{2}h, \; x_i + \frac{1}{2}h)$, při čemž $x_i + \frac{1}{2}h = x_{i+1} - \frac{1}{2}h$. Zařazování prvků do tříd se může provádět trojím způsobem: do intervalu spadají hodnoty znaku x splňující nerovnosti

1. $x_i - \frac{1}{2}h < x \leqq x_i + \frac{1}{2}h$,

2. $x_i - \frac{1}{2}h \leqq x < x_i + \frac{1}{2}h$,

3. $x_i - \frac{1}{2}h < x < x_i + \frac{1}{2}h$, při čemž tam bude zařazena polovina prvků se znakem $x_i - \frac{1}{2}h$ a polovina prvků se znakem $x_i + \frac{1}{2}h$. Není-li stanovena dolní hranice prvního intervalu, zařadí se tam všechny hodnoty proměnné $x < x_1 + \frac{1}{2}h$, po případě $x \leqq x_1 + \frac{1}{2}h$. Do posledního intervalu neohraničeného shora jsou pak zahrnuty hodnoty proměnné $x \geqq x_l - \frac{1}{2}h$, po případě $x > x_l - \frac{1}{2}h$.

Hodnoty spojité proměnné mohou býti měřeny jen s určitou přesností, takže také v tomto případě vycházejí ze statistického šetření jednotlivé hodnoty isolované. Je důležito, aby stanovené hranice ukazovaly, na které desetinné místo bylo měřeno [5]. Toho docilujeme buď tím, že jsou

v hranicích přímo vyznačeny krajní hodnoty, které do intervalu spadají, nebo hranice vymezují hodnoty spadající do intervalu pomocí dalšího desetinného místa.

Objasníme si to příkladem. Představme si, že byl stanoven znak u každého prvku souboru na setiny určité jednotky míry (cm, kg, ...) a hned zaokrouhlován na desetiny; to znamená, že dostaneme výsledky, o nichž říkáme, že byly měřeny s přesností na desetiny.

Zaokrouhlení bylo prováděno třeba podle dohody, že zlomky 0,01 až 0,04 se zanedbají, 0,06 až 0,09 dávají 0,1 a 0,05 u sudého čísla na předchozím desetinném místě dává 0,1, u lichého se zanedbá.

Kdybychom tvořili na př. intervaly délky 0,5, tedy po pěti hodnotách znaku (počet lichý) dostali bychom třeba 80,0—80,4, 80,5—80,9, ... a střed intervalu čili třídní znak bude 80,2, 80,7, ...

Kdybychom tvořili intervaly délky 1,0, tedy po desíti (sudý počet), byl by střed intervalu a tedy znak 80,45.

Mohli bychom však při měření s přesností na desetiny vyznačit hranice v prvním případě čísly 79,95, 80,45, 80,95, ... což také určuje jednoznačně, že v prvém intervalu jsou hodnoty 80,0, 80,1, 80,2, 80,3, 80,4 a podobně ve druhém. Tím je současně vyjádřeno, že z hodnot stanovených na setiny spadají do prvního intervalu 79,96—80,44 a do druhého 80,45—80,95, z čehož také vyplývá střed 80,2, 80,7.

V druhém případě pak spadají do intervalu hodnoty 79,96—80,95, takže střed je 80,45, neboť 5 na dalším místě se zanedbává, když předchází liché číslo.

(3,6) Sestrojení tabulky skupinového rozdělení četností pro daný příklad. Sestrojme nyní skupinové rozdělení četností pro materiál našeho příkladu. Znak byl

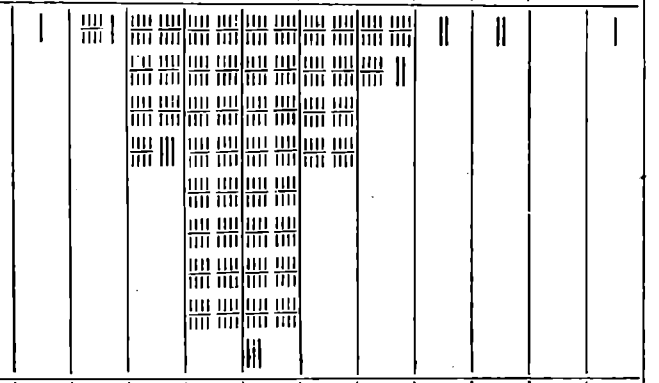
měřen na celé jednotky. Stanovíme hranice na př. druhým způsobem pomocí dalšího desetinného místa. Zvolíme interval délky $h = 15$. Dolní hranice prvního intervalu pak bude $x_1 - \frac{1}{2}h = 22{,}5$. Dostaneme tak tabulku o 12 třídách, kde v prvním sloupci uvedeme hranice intervalů, v druhém průměr hodnot proměnné spadajících do intervalu čili třídní znak a ve třetím sloupci příslušnou pozorovanou četnost třídní.

Tabulka 1.

Třídní hranice	Třídní znak	Četnost	Kumulativní četnost	Relativní četnost	Kumulativní relativní četnost
1	2	3	4	5	6
22,5					
37,5	30	1	1	0,4	0,4
52,5	45	1	2	0,4	0,8
67,5	60	6	8	2,2	3,0
82,5	75	38	46	14,1	17,1
97,5	90	80	126	29,6	46,7
112,5	105	83	209	30,7	77,4
127,5	120	39	248	14,5	91,9
142,5	135	17	265	6,3	98,2
157,5	150	2	267	0,7	98,9
172,5	165	2	269	0,7	99,6
187,5	180	0	269	0,0	99,6
202,5	195	1	270	0,4	100,0
Celkem		270		100,0	

$r = 270$

Vzhledem k poměrně malému rozsahu souboru lze provésti roztřídění do skupin stanovených třídními intervaly buď metodou skládání lístků nebo metodou čárkovací, kterou si znázorníme takto:

22,5 až 37,5	37,5 až 52,5	52,5 až 67,5	67,5 až 82,5	82,5 až 97,5	97,5 až 112,5	112,5 až 127,5	127,5 až 142,5	142,5 až 157,5	157,5 až 172,5	172,5 až 187,5	187,5 až 202,5
											
1	1	6	38	80	83	39	17	2	2	0	1

K některým účelům je důležito znáti počet prvků, jejichž hodnota znaku nepřevyšuje hranici určitého intervalu třídního; ten počet udává kumulativní čili součtové rozdělení četností, které dostáváme z rozdělení četností (sloupec 3) zapsáním četnosti první třídy a postupným načítáváním jednotlivých třídních četností k předchozímu součtu (sloupec 4) $s_i = s_{i-1} + n_i$, $s_1 = n_1$.

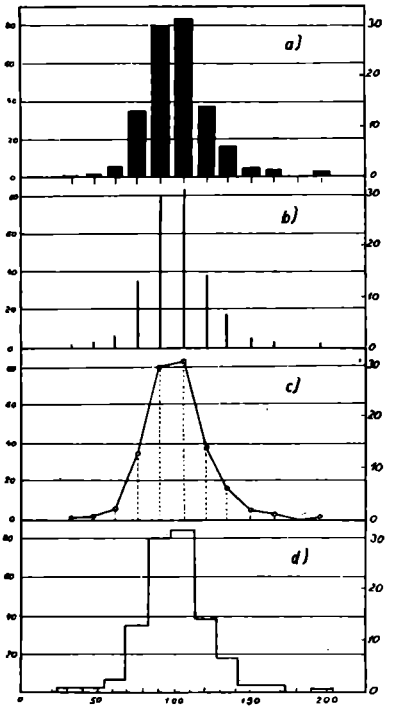
Tato tabulka nám udává na př. ihned počet prvků, jejichž hodnota znaku $x \leqq x_i + \frac{1}{2}h$ a tedy snadno pak počet ostatních, jejichž hodnota znaku je větší čili četnosti třídění dichotomického. V našem případu, se jedná o znak kvantitativní; při něm dichotomické třídění znamená, že buď se vyskytují jeho hodnoty až do $x_i + \frac{1}{2}h$ nebo se nevyskytují. Je to obdoba alternativy při znaku kvalitativním (muž — žena). Třídíme-li podle mnoha stupňů znaku, na-

zýváme takové třídění polytomické. Kromě napozorované četnosti prvků, t. zv. absolutní četnosti třídní, má v dalších úvahách velký význam relativní četnost třídní, která je podílem absolutní třídní četnosti a celkového rozsahu souboru $f_i = \frac{n_i}{r}$. Velmi často se uvádějí relativní četnosti v procentech (sloupec 5).

Tabulka relativních četností, jejichž součet se rovná jedné $\sum_{i=1}^{l} f_i = 1$ je jen tehdy úplná, je-li současně uveden rozsah souboru r. Z ní se zase odvozuje kumulativní rozdělení relativních četností (sloupec 6) $F_i = F_{i-1} + f_i$.

Seznámili jsme se s hlavními metodami zhuštěného podávání výsledků jednak pomocí charakteristik, jednak formou tabelární.

(3,7) Grafické podávání statistických výsledků. K dalšímu usnadnění přehledné a jasné představy o studovaném souboru užívá statistika grafického podávání výsledků šetření. Vytváří grafické profily souboru tím, že znázorňuje jeho rozdělení četností podle vyšetřovaných znaků.

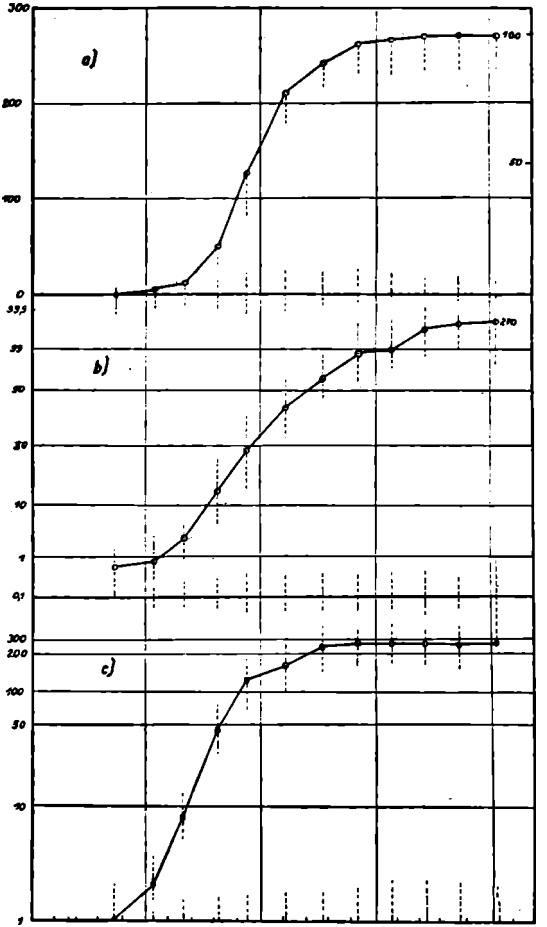


Obr. 3. Grafické znázornění rozdělení četností z tab. 1.

Používá se k tomu účelu v systému pravoúhlých souřadnic vodorovné osy úseček pro stupnici hodnot statistické proměnné a osy pořadnic pro četnosti. Znázornění lze pak provésti několika způsoby.

Diagram tyčkový vzniká vztyčením řady pruhů výšky rovnající se třídní četnosti, jejichž střed je v prostředních bodech tříd, t. j. v bodech znázorňujících třídní znak (obr. 3a). Šířka pruhů bývá různá; zvláště často se užívá úzkých úseček délky rovné třídní četnosti (obr. 3b) vztyčených v bodech třídních znaků. Velmi vhodným prostředkem grafického znázornění je mnohoúhelník (polygon) četností (obr. 3c), který dostaneme, jestliže tečky nebo kroužky ve výši třídních četností nad body třídních znaků spojíme úsečkami. Krajové tečky se spojí se středem nejbližšího intervalu na ose úseček, čímž je polygon uzavřen. Oblíbeným je znázornění pomocí histogramu četností čili sloupkový diagram sestávající z obdélníků, jejichž základna se rovná intervalu třídnímu a výška třídní četnosti, dělené délkou intervalu (obr. 3d). V histogramu představují plochy obdélníků třídní četnosti; je to jako bychom měřili stupnice třídním intervalem jako jednotkou. Kraje sloupků představují třídní hranice. V obraze a) až d) je současně patrno, že znázorňují také relativní četnosti uvedené na pravé straně obrazu, takže bývá výhodné užíti obou stupnic. Jako při tabulce, tak i při diagramu relativních četností nemá scházeti uvedení rozsahu souboru.

Znázorníme-li způsobem c) data sloupce 4. nebo 6., ale pro horní hranice příslušného intervalu, dostaneme kumulativní diagram četnosti neboli ogiv (obr. 4a), kde je vyznačena na levo aritmetická stupnice pro četnosti absolutní a napravo pro četnosti relativní, vyjádřené v procentech. Užívá se s prospěchem pro relativní četnosti také stupnice nomografické (pravděpodobnostní, obr. 4b), která převádí součtovou křivku Gaussovu (str. 84) na přímku [9], [7]. K výkladu stupnice a účelnosti její můžeme přistoupit až později.



Obr. 4. Grafické znázornění kumulativní četnosti. (Součtové křivky.)
a) V aritmetické stupnici pro absolutní i relativní četnosti.
b) V pravděpodobnostní stupnici pro relativní četnosti.
c) V logaritmické stupnici pro absolutní i relativní četnosti.

Stupnice logaritmické se užívá tam, kde je třeba diagramu méně citlivého na malé variace, kde by tedy v citlivém diagramu nevynikl celkový hlavní průběh (obr. 4c).

Úloha: Znázorněte pomocí histogramu a) úmrtnost mužů na rakovinu a b) úmrtnost obojího pohlaví na chřipku, která je uvedena v počtu případů na 10 000 žijících v letech věkové stupnice s nestejnými intervaly. Četnost v první věkové třídě se vztahuje na 10 000 živě narozených. V případě a) dostanete tak zv. J-křivku a v případě b) tak zv. U-křivku.

Věková třída	a)	b)
0—	0,1	13,7
1—	0,1	2,3
5—	0,02	0,6
15—	0,2	0,7
30—	9,4	1,5
60—	65,0	7,8
70—	100,9	19.2

Dalšího zhuštění pozorovaného materiálu dosahujeme spojením obou dříve naznačených cest, které vyžaduje, abychom upravili způsob výpočtu momentových charakteristik pro skupinové rozdělení četností.

(3,8) Základní charakteristiky a jejich výpočet pro skupinové rozdělení četností. Je-li z celé posloupnosti r hodnot x_i jen l hodnot od sebe různých, sestavujeme jednoduchou tabulku rozdělení četností

$$\frac{x_1, x_2, \ldots, x_l}{n_1, n_2, \ldots, n_l},$$

pro niž pak obecné momenty jsou vyjádřeny rovnicí

$$\mu'_{x,k} = \frac{1}{r}(n_1x_1^k + n_2x_2^k + \dots + n_lx_l^k) = \frac{1}{r}\sum_{i=1}^{l} n_ix_i^k, \quad (13)$$

při čemž $n_1 + n_2 + \dots + n_l = r$. Vidíme totiž, že součty $\sum_{i=1}^{l} n_ix_i^k$ jsou numericky ekvivalentní součtům $\sum_{i=1}^{r} x_i^k$. Na př. pro $k = 1$ bude $\sum_{i=1}^{l} x_in_i = \sum_{i=1}^{r} x_i$ pro $k = 2$ je $\sum_{i=1}^{l} x_i^2n_i = \sum_{i=1}^{r} x_i^2$ atd. V součtech $\sum_{i=1}^{l} x_i^kn_i$ jsou již sečteny určité skupiny hodnot x^k, a to těch, které jsou stejné.

Pro momenty kolem aritmetického průměru platí totéž co jsme dříve odvodili a rovněž vztahy mezi nimi a momenty obecnými se nemění.

(3,9) Výpočet momentů metodou vhodně zvoleného počátku. Přistoupíme nyní k výpočtu momentů pro skupinové rozdělení četností. Výpočet se zjednoduší zpravidla metodou vhodně zvoleného počátku nebo metodou součtovou. Provedeme jej nejprve pro náš numerický příklad první metodou.

Máme l tříd četnosti a tedy l třídních znaků, které zastupují všechny hodnoty proměnné. Četnost n_i náleží celé třídě, v níž všechny hodnoty proměnné jsou zastoupeny třídním znakem x_i. Je tudíž obecný moment dán výrazem

$$\mu'_{x,k} = \frac{1}{r}\sum_{i=1}^{l} x_i^kn_i.$$

Zjednodušení výpočtu dosáhneme tím, že zavedeme novou proměnnou, pro niž zvolíme nový vhodný počátek x_0 a budeme ji měřit délkou intervalu jako novou jednotkou. Celý výpočet tedy bude proveden v jednotce h, v délce třídního intervalu. Nová proměnná bude

$$u_i = \frac{x_i - x_0}{h} = \frac{\eta_i}{h}, \text{ takže } x_i = hu_i + x_0. \quad (14)$$

Aritmetický průměr se vyjádří takto

$$\bar{x} = \frac{1}{r}\sum_{i=1}^{l} x_i n_i = \frac{1}{r}\sum_{i=1}^{l} (hu_i + x_0)\, n_i =$$
$$= \frac{1}{r}\left[h\sum_{i=1}^{l} u_i n_i + x_0 \sum_{i=1}^{l} n_i\right],$$

z čehož je patrno, že

$$\bar{x} = h\bar{u} + x_0. \tag{15}$$

Druhý moment kolem aritmetického průměru pro proměnnou u_i odvodíme podobně.

Druhý obecný moment pro proměnnou $\eta_i = x_i - x_0$ je

$$\mu'_{\eta,2} = \frac{1}{r}\sum_{i=1}^{l} \eta_i^2 n_i = \frac{1}{r} h^2 \sum_{i=1}^{l} u_i^2 n_i.$$

Víme z rovnice (5), že

$$\mu_{\eta,2} = \mu'_{\eta,2} - \bar{\eta}^2$$

a vzhledem k základnímu teorému o momentech $\mu_{x,k} = \mu_{\eta,k}$ tedy také

$$\mu_{x,2} = \mu'_{\eta,2} - \bar{\eta}^2 = h^2\left[\frac{1}{r}\sum_{i=1}^{l} u_i^2 n_i - \left(\frac{1}{r}\sum_{i=1}^{l} u_i n_i\right)^2\right]$$

čili

$$\mu_{x,2} = h^2[\mu'_{u,2} - \bar{u}^2],$$

což lze psáti také

$$\mu_{x,2} = h^2 \mu_{u,2} \tag{16}$$

a směrodatná odchylka bude tudíž vyjádřena rovnicí

$$\sigma_x = h\sigma_u. \tag{17}$$

Pro třetí moment dostaneme

$$\mu'_{\eta,3} = \frac{1}{r}\sum_{i=1}^{l} \eta_i^3 n_i = \frac{1}{r} h^3 \sum_{i=1}^{l} u_i^3 n_i,$$
$$\mu_{\eta,3} = \mu'_{\eta,3} - 3\bar{\eta}\mu'_{\eta,2} + 2\bar{\eta}^3.$$

a podle základního teorému o momentech

$$\mu_{x,3} = \mu_{\eta,3} =$$

$$= h^3 \left[\frac{1}{r} \sum_{i=1}^{l} u_i^3 n_i - 3 \left(\frac{1}{r} \sum_{i=1}^{l} u_i n_i \right) \left(\frac{1}{r} \sum_{i=1}^{l} u_i^2 n_i \right) + 2\bar{u}^3 \right]$$

$$\mu_{x,3} = h^3 [\mu'_{u,3} - 3\bar{u}\mu'_{u,2} + 2\bar{u}^3]$$

čili

$$\mu_{x,3} = h^3 \mu_{u,3}. \tag{18}$$

Vzhledem k tomu je patrno, že šikmost nebo kosost se touto změnou proměnné nemění, neboť

$$\alpha_{x,3} = \frac{\mu_{x,3}}{\sigma_x^3} = \frac{\mu_{u,3}}{\sigma_u^3} = \alpha_{u,3}. \tag{19}$$

Stejným způsobem si čtenář ukáže, že platí pro čtvrté momenty

$$\mu_{x,4} = h^4 \mu_{u,4} \text{ a tudíž } \alpha_{x,4} = \alpha_{u,4}. \tag{20}$$

Obecné odvození pro k-tý moment nečiní potíží.

Kontrola numerického výpočtu momentů proměnné u se provádí t. zv. metodou posunutých momentů čili Charlierovým testem. Tento postup se zakládá na binomické větě

$$(u_i + 1)^3 = u_i^3 + 3u_i^2 + 3u_i + 1,$$

takže vynásobíme-li třídní četností n_i a sečteme pro všechna i dostaneme

$$\sum_{i=1}^{l} (u_i + 1)^3 n_i = \sum_{i=1}^{i} u_i^3 n_i + 3 \sum_{i=1}^{l} u_i^2 n_i + 3 \sum_{i=1}^{l} u_i n_i + \sum_{i=1}^{l} n_i. \tag{21}$$

Počítáme-li momenty až do čtvrtého řádu, pak provádíme kontrolu podle $(u_i + 1)^4$.

(3,10) Výpočet momentů metodou součtovou. Také při této metodě zvolíme pomocný počátek na příklad tak, že první hodnotě znaku, pro kterou se v tabulce vyskytuje nějaká četnost, přidělíme znak $u = 1$, znak další třídy označíme $u = 2$ atd. Sčítáme pak četnosti zdola přes celou

tabulku a pro každou třídu vyznačíme příslušný mezisoučet. Tento součtový sloupec pak znovu sečítáme zdola a to opakujeme tolikrát, kolik momentů potřebujeme. Poslední součet v každém sloupci označíme postupně $S_0, S_1, S_2, \ldots, S_k \ldots$ Můžeme si odvoditi, že

$$S_0 = \sum i\, n_i = N, \quad S_1 = \sum i\, in_i,$$

$$S_2 = \sum i \frac{i\,(i+1)}{2!} n_i, \ \ldots, \ S_k = \sum i \binom{i+k-1}{k} n_i, \ \ldots$$

Označíme-li $s_k = \dfrac{S_k}{S_0}$ a tyto hodnoty s_k vyjádříme pomocí momentů proměnné u kolem počátku $u = 0$, dostaneme

$$s_1 = \mu'_{u,1} = \bar{u}$$
$$s_2 = \tfrac{1}{2}(\mu'_{u,2} + \bar{u})$$
$$s_3 = \tfrac{1}{6}(\mu'_{u,3} + 3\mu'_{u,2} + 2\bar{u})$$
$$s_4 = \tfrac{1}{24}(\mu'_{u,4} + 6\mu'_{u,3} + 11\mu'_{u,2} + 6\bar{u}).$$

Z těchto hodnot pak plynou vzorce pro obecné momenty proměnné u

$$\bar{u} = \mu'_{u,1} = s_1$$
$$\mu'_{u,2} = 2s_2 - s_1$$
$$\mu'_{u,3} = 6s_3 - 6s_2 + s_1$$
$$\mu'_{u,4} = 24s_4 - 36s_3 + 14s_2 - s_1.$$

Momenty kolem aritmetického průměru určíme obvyklým způsobem dříve uvedeným. Výrazy pro výpočet momentů kolem aritmetického průměru přímo z hodnot s_k jsou dosti složité a proto je neuvádíme.

Součtové metody se méně používá, protože pracuje s velkými čísly, zvláště při větším počtu tříd a vyšších momentech. Početní postup při jejím použití je nejlepé patrný z příkladu podle tab. 2.

x_i	u_i	n_i	$\Sigma(3)$	$\Sigma(4)$	$\Sigma(5)$	$\Sigma(6)$
(1)	(2)	3	4	5	6	7
30	1	1	270	1530	5345	14723
45	2	1	269	1260	3815	9378
60	3	6	268	991	2555	5563
75	4	38	262	723	1564	3008
90	5	80	224	461	841	1444
105	6	83	144	237	380	603
120	7	39	61	93	143	223
135	8	17	22	32	50	80
150	9	2	5	10	18	30
165	10	2	3	5	8	12
180	11	—	1	2	3	4
195	12	1	1	1	1	1
Σ		270	1530	5345	14723	35069
		S_0	S_1	S_2	S_3	S_4

$S_0 = 270$

$S_1 = 1530$ $\qquad s_1 = \dfrac{S_1}{S_0} = 5{,}6667$

$S_2 = 5345$ $\qquad s_2 = \dfrac{S_2}{S_0} = 19{,}7963$

$S_3 = 14723$ $\qquad s_3 = \dfrac{S_3}{S_0} = 54{,}5296$

$S_4 = 35069$ $\qquad s_4 = \dfrac{S_4}{S_0} = 129{,}8852$

$\mu'_{u,1} = \bar{u} = s_1 = 5{,}667$

$\mu'_{u,2} = 2s_2 - s_1 = 33{,}926$

$\mu'_{u,3} = 6s_3 - 6s_2 + s_1 = 214{,}067$

$\mu'_{u,4} = 24s_4 - 36s_3 + 14s_2 - s_1 = 1425{,}661$

$$\mu_{u,2} = \mu'_{u,2} - \bar{u}^2 = 1{,}815$$
$$\mu_{u,3} = \mu'_{u,3} - 3\mu'_{u,2}\bar{u} + 2\bar{u}^3 = 1{,}250$$
$$\mu_{u,4} = \mu'_{u,4} - 4\mu'_{u,3}\bar{u} + 6\mu'_{u,2}\bar{u}^2 - 3\bar{u}^4 = 16{,}456$$

Je možno voliti počátek na př. poblíž třídy s největší četností, takže se dosáhne dvojími součty nahoru a dolů menších čísel, ale výrazy pro momenty jsou zase trochu složitější.

(3,11) Opravy momentů. Tím, že při skupinovém rozdělení četností zastupuje prostřední hodnota třídního intervalu všechny hodnoty znaku dotyčného intervalu, vzniká při výpočtu momentů jistá odchylka (chyba) od momentů, které by byly počítány ze všech hodnot znaku, jak byly napozorovány nebo pro spojitou proměnnou jako funkcionální momenty

$$m_{x,k} = \frac{1}{r}\int_a^l (x - \bar{x})^k \, n(x)\, \mathrm{d}x.$$

Proto se momenty, vypočítané svrchu uvedeným postupem, opravují t. zv. Sheppardovou korekcí, takže pro opravené momenty platí rovnice

$$_0\mu_{x,2} = \mu_{x,2} - \tfrac{1}{12}h^2, \qquad (22)$$
$$_0\mu_{x,3} = \mu_{x,3}, \qquad (23)$$
$$_0\mu_{x,4} = \mu_{x,4} - \tfrac{1}{2}h^2\mu_{x,2} + \tfrac{7}{240}h^4 \qquad (24)$$

a je-li délka třídního intervalu rovna jedné, položíme $h = 1$. Tak tedy v případě proměnné u, kde je délka intervalu zvolena za jednotku, bude na př.

$$_0\mu_{u,2} = \mu_{u,2} - \tfrac{1}{12}.$$

Poněvadž

$$\mu_{x,2} = h^2\mu_{u,2}, \text{ bude tedy } {}_0\mu_{x,2} = h^2\,{}_0\mu_{u,2}.$$

(3,12) Schema výpočtu. Je účelno zachovávati při výpočtu momentů určitý pořádek v zapisování výsledků; provedeme tedy podrobný výpočet pro náš numerický příklad.

Tabulka 2.

Třídní hranice	Třídní znak	Četnost n_i	u_i	$u_i n_i$	$u_i^2 n_i$	$u_i^3 n_i$	$u_i^4 n_i$	$(u_i+1)^4 n_i$
22,5	30	1	−5	− 5	25	−125	625	256
37,5	45	1	−4	− 4	16	− 64	256	81
52,5	60	6	−3	−18	54	−162	486	96
67,5	75	38	−2	−76	152	−304	608	38
82,5	90	80	−1	−80	80	− 80	80	0
97,5	105	83	0	0	0	0	0	83
112,5	120	39	1	39	39	39	39	624
127,5	135	17	2	34	68	136	272	1377
142,5	150	2	3	6	18	54	162	512
157,5	165	2	4	8	32	128	512	1250
172,5	180	0	5	0	0	0	0	0
187,5	195	1	6	6	36	216	1296	2401
Součty ...		270		−183 + 93		−735 +573		
Celkem ...				− 90	520	−162	4336	6718

$$x_0 = 105, h = 15.$$

O správnosti výpočtu se přesvědčíme testem Charlierovým, neboť jsme si k tomu cíli připravili poslední sloupec tabulky.

$$4336 - 4 \times 162 + 6 \times 520 - 4 \times 90 + 270 = 6718,$$

což je provedeno podle součtu rovnic

$$(u_i + 1)^4 n_i = u_i^4 n_i + 4u_i^3 n_i + 6u_i^2 n_i + 4u_i n_i + n_i$$

pro $i = 1, 2, \ldots, l$.

Momenty pomocné proměnné u, obecné:

$$\mu'_{u,1} = \bar{u} = -0{;}3333$$
$$\mu'_{u,2} = 1{;}9259 \qquad \bar{u}^2 = 0{,}1111$$
$$\mu'_{u,3} = -0{;}6000 \qquad -3\bar{u}\mu'_{u,2} = +1{,}9257$$
$$\mu'_{u,4} = 16{;}0593 \qquad 2\bar{u}^3 = -0{,}0741$$

$$-4\bar{u}\mu'_{u,3} = -0{,}7999$$
$$+6\bar{u}^2\mu'_{u,2} = +1{,}2838$$
$$-3\bar{u}^4 = -0{,}0370$$

kolem průměru $\bar{u}$:

$$\mu_{u,2} = 1{,}8148 \qquad \sigma_u = 1{,}3471$$
$$\mu_{u,3} = 1{,}2534 \qquad \sigma_u\mu_{u,2} = 2{,}4447$$
$$\mu_{u,4} = 16{,}5062 \qquad \mu_{u,2}{}^2 = 3{,}2935$$

$$_0\mu_{u,2} = 1{,}7315 \qquad {}_0\sigma_u = 1{,}3159$$
$$_0\mu_{u,3} = 1{,}2534 \qquad {}_0\sigma_u\,{}_0\mu_{u,2} = 2{,}2785$$
$$_0\mu_{u,4} = 15{,}6280 \qquad {}_0\mu_{u,2}{}^2 = 2{,}9981$$

Momenty proměnné x kolem průměru

$$\bar{x} = 100{,}00$$

$$\mu_{x,2} = 408{,}33 \quad \sigma_x = 20{,}21 \quad {}_0\mu_{x,2} = 389{,}59 \quad {}_0\sigma_x = 19{,}74$$
$$\mu_{x,3} = 4\,230 \quad \alpha_{x,3} = 0{,}51 \quad {}_0\mu_{x,3} = 4\,230 \quad {}_0\alpha_{x,3} = 0{,}55$$
$$\mu_{x,4} = 835\,600 \quad \alpha_{x,4} = 5{,}01 \quad {}_0\mu_{x,4} = 791\,200 \quad {}_0\alpha_{x,4} = 5{,}21$$

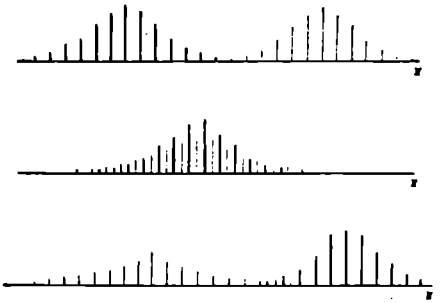
O rozdělení četností nabýváme pomocí charakteristik jisté přibližné představy. Tak průměr je charakteristikou polohy souboru na stupnici hodnot znaku, směrodatná odchylka nebo její čtverec je výrazem rozptylu, šikmost nebo kosost udává míru nesouměrnosti rozdělení četností. Na obr. 5 jsou znázorněna dvě symetrická rozdělení četnosti

a) s týmž rozptylem, ale různou polohou,

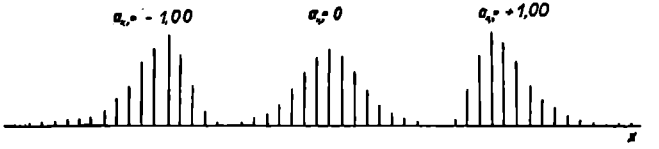
b) s různými rozptyly a touž polohou,

c) s různými rozptyly a různou polohou.

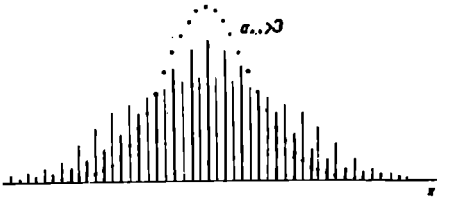
Obr. 6 ukazuje souměrné rozdělení jakož i zápornou a kladnou šikmost. Konečně obr. 7 osvětluje kladnou a zápornou špičatost (exces), která se měří srovnáváním s nor-



Obr. 5. Dvě základní charakteristiky rozdělení četností. (Poloha — Rozptyl.)



Obr. 6. Třetí charakteristika rozdělení četnosti. (Šikmost pro tři zvláštní hodnoty.)



Obr. 7. Čtvrtá charakteristika rozdělení četností. (Exces kladný a záporný přirovnán k normálnímu $\alpha_{x_i 4} = 3$.)

mální křivkou Laplace-Gaussovou (viz 82), jejíž špičatost je $\alpha_{r,4} - 3 = 0$ (úsečky silněji vytažené).

(3,13) Přesnost průměru a směrodatné odchylky. Otázku po přesnosti, které docílíme pro průměr a směrodatnou odchylku, můžeme zodpověděti tak, že stačí obyčejně uváděti to desetinné místo, na které jsou měřeny hodnoty znaku.

Často se také postupuje dále podle směrnice, že počet desetinných míst v numerických hodnotách charakteristik se řídí podle směrodatné odchylky. Vyjádří se tedy směrodatná odchylka třemi, resp. dvěma významnými číslicemi a charakteristiku uvedeme na týž počet desetinných míst jako směrodatnou odchylku nebo o jedno místo méně. Při tom se doporučuje dbáti, aby zaokrouhlení vzniklé vynecháním desetinných míst nepřekročilo 0,1 směrodatné odchylky.

Jinak bývá často pro usnadnění práce třeba zaokrouhlovat hodnoty znaku. Pak je nutno odhadovat největší možné hranice účinku zaokrouhlení na výsledek. Takový odhad lze prováděti vhodně na př. takto: Nahradíme-li přesné číslo x číslem x', je absolutní chyba $\vartheta = x - x'$, takže je přesné číslo $x = x' + \vartheta$, což můžeme také psáti $x = x'\left(1 + \frac{\vartheta}{x'}\right)$. Potom zlomek $\frac{\vartheta}{x'}$ se označuje symbolem ε nebo e a nazývá se relativní chybou čísla x'. Přesné číslo je pak vyjádřeno $x = x'(1 + \varepsilon)$. Máme nyní dvě čísla x' a y', jejichž relativní chyby jsou ε resp. e, pak dostáváme relativní chybu součtu a rozdílu jejich

$$x \pm y = x'(1 + \varepsilon) \pm y'(1 + e) = x' \pm y' + x'\varepsilon \pm y'e = \\ = (x' \pm y')\left(1 + \frac{x'\varepsilon \pm y'e}{x' \pm y'}\right). \qquad (25)$$

Můžeme-li považovati relativní chyby obou členů za stejné, dostaneme jednoduše $x \pm y = (x' \pm y')(1 + \varepsilon)$, takže relativní chyba součtu nebo rozdílu se rovná relativní chybě

jeho členů. Tento výsledek a ovšem také rovnici (25) lze snadno rozšířiti na libovolný počet členů.

Jsou-li chyby ε a e malé, dostaneme provedením naznačených operací přibližné vyjádření chyby

součinu $x \,.\, y = x' \,.\, y' \,(1 + \varepsilon)\,(1 + e) \doteq x'y' \,(1 + \varepsilon + e)$

mocniny $x^m = [x' \,(1 + \varepsilon)]^m \doteq x'^m \,(1 + m\varepsilon)$

podílu $\dfrac{x}{y} = \dfrac{x'}{y'} \dfrac{1 + \varepsilon}{1 + e} \doteq \dfrac{x'}{y'} \,(1 + \varepsilon - e).$

Uvažujeme-li o chybě průměru vzniklé zaokrouhlením hodnot znaku x_i, které jsou nahrazeny hodnotami x'_i, při čemž může býti ε největší relativní chyba ze zaokrouhlení a tu vezmeme pro všechny hodnoty, pak $x'_i \,(1 - \varepsilon) \leqq x_i \leqq x'_i \,(1 + \varepsilon)$. Utvoříme-li ze všech tří hodnot průměry vidíme, že $\bar{x}' \,(1 - \varepsilon) \leqq \bar{x} \leqq \bar{x}' \,(1 + \varepsilon)$, čili hranice chyb průměru zaokrouhlených čísel nepřekročí největší možnou chybu vzniklou zaokrouhlením jednotlivých hodnot znaku. Je zajímavo, že vypočítáme-li si pro náš příklad průměr z hodnot seřazených na str. 19, dostaneme $\bar{x} = 92{,}5$, kdežto průměr z hodnot zastoupených třídními znaky v tabulce č. 2 je 100,0. Poněvadž největší možná chyba je 7,5 vidíme, že je tato hodnota průměru právě na hranici možných chyb vzniklých seskupením do tříd, ač velmi často se chyby značně kompensují.

(3,14) Přehled charakteristik. Vzhledem k tomu, že kromě momentových charakteristik se užívá často k některým účelům také jiných, seznámíme se s těmi nejdůležitějšími.

1. Charakteristiky polohy. a) Aritmetický průměr je nejrozšířenější charakteristikou polohy nebo také mírou ústřední tendence. Jeho prvá podstatná vlastnost je, že součet odchylek hodnot znaku u všech jednotek souboru od aritmetického průměru se rovná nule. $\mu_{x,1} = 0$. Druhá vlastnost je, že součet čtverců těchto odchylek je minimem. Známe-li průměry jednotlivých částí souboru, násobíme je

jejich rozsahem a dělíme rozsahem celého souboru, abychom dostali průměr souboru. Zvětší-li se nebo zmenší-li se všechny hodnoty znaku o konstantu, zvětší nebo zmenší se o ni také průměr.

$$\mu'_{x\pm a,1} = \frac{1}{r}(x_1 \pm a + x_2 \pm a + \ldots + x_r \pm a) =$$

$$= \frac{1}{r}\sum_{i=1}^{r} x_i \pm a = \mu'_{x,1} \pm a.$$

Násobí-li se hodnoty znaku konstantou, je průměr násoben touž konstantou

$$\mu'_{ax,1} = \frac{1}{r}(ax_1 + ax_2 + \ldots + ax_r) = a\frac{1}{r}\sum_{i=1}^{r} x_i = a\mu'_{x,1}.$$

b) Poznali jsme již také medián, jehož se zvláště užívá tehdy, když jsou hodnoty znaku nesnadno měřitelné, ale lze je snadno aspoň seřadit podle velikosti. Nezávisí na krajových hodnotách znaku, které mají na př. značný vliv na průměr aritmetický. Můžeme jej určit i když jsou intervaly nekonečně velké, takže znemožňují výpočet průměru. Řekneme-li, že medián je velký, víme, že polovina pozorovaná je jistě velká, kdežto u aritmetického průměru nemůžeme ničeho říci o celém množství pozorování, neboť jeho vysoká hodnota může býti způsobena několika isolovanými případy. Proto se často dává mediánu přednost před průměrem na př. při statistice mezd. Stanovení mediánu ze skupinového rozdělení četností provedeme pro náš příklad. Ze sloupce kumulativní četnosti vidíme, že prostřední dva členy (při sudém rozsahu) jsou v intervalu 97,5 až 112,5 mezi 135. a 136. pozorováním čili mezi 9. a 10. prvkem intervalu. Budeme předpokládati, že hodnoty proměnné jsou v intervalu stejnoměrně rozloženy, takže délku intervalu s četností 83 rozdělíme úměrou a dostaneme pro devátou hodnotu

$$97{,}5 + 15\,\tfrac{9}{83} = 99{,}13 \text{ a pro desátou } 99{,}31.$$

Mezi těmito dvěma čísly leží medián, za nějž vezmeme jejich střed, takže $\tilde{x} = 99{,}22$.

Můžeme psáti obecně výraz pro medián, který má býti v intervalu $x_m - \frac{1}{2}h$, $x_m + \frac{1}{2}h$ při lichém rozsahu souboru r

$$x = x_m - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r+1}{2} - s_{m-1}\right\} \tag{26}$$

a při sudém r je mezi dvěma výrazy

$$x_m - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r}{2} - s_{m-1}\right\} < \tilde{x} < x_m - \\ - \frac{h}{2} + \frac{h}{n_m}\left\{\frac{r}{2} + 1 - s_{m-1}\right\} \tag{27}$$

jak čtenář snadno sám nahlédne.

c) Modus je nejčetnější hodnota znaku. Z rozdělení četností, které zachovává jednotlivé pozorované hodnoty proměnné, se určí modus jednoduše jako hodnota, jíž přísluší největší četnost n_a. Obtíž vzniká, máme-li určiti tuto hodnotu pro skupinové rozdělení četností, které se nejčastěji v praksi vyskytuje. Udávati střed intervalu s největší třídní četností by mělo malý význam, neboť ten závisí na volbě stupnice pro třídní intervaly. Proto se určuje obyčejně modus přibližně jako hodnota, která přísluší maximu křivky proložené co nejtěsněji skutečným rozdělením četností. V případech mírně nesouměrných rozdělení četností si pomáháme proložením paraboly druhého stupně $y = c_2x^2 + c_1x + c_0$ bodem znázorňujícím největší četnost přiřazenou jejímu třídnímu znaku a obdobným bodem sousedním s každé strany. Položíme-li počátek do třídního znaku třídy s největší četností, dostaneme k určení konstant paraboly tři rovnice [10] [11]

$$\begin{aligned} n_{a-h} &= c_2h^2 - c_1h + c_0 \\ n_a &= c_0 \\ n_{a+h} &= c_2h^2 + c_1h + c_0, \end{aligned}$$

takže řešením dostáváme

$$n_{a-h} + n_{a+h} - 2n_a = 2c_2h^2$$
$$n_{a+h} - n_{a-h} = 2c_1h$$

Modus, jakožto úsečka vrcholu paraboly je stanoven podmínkou, že prvá derivace se rovná nule $y' = 2c_2x + c_1 = 0$ čili $x' = -\frac{c_1}{2c_2}$, takže dosazením z rovnic dostáváme

$$x' = -\frac{h}{2}\,\frac{n_{a+h} - n_{a-h}}{n_{a+h} - 2n_a + n_{a-h}},$$

což je tedy poloha modu měřená od počátku v třídním znaku intervalu s maximální četností, takže modus $\hat{x}$ bude pak vyjádřen vztahem

$$\hat{x} = x_a - \frac{h}{2}\,\frac{n_{a+h} - n_{a-h}}{n_{a+h} - 2n_a + n_{a-h}}. \qquad (28)$$

Pro náš numerický příklad skupinového rozdělení četností z toho vyplývá toto vyjádření $x_a = 105$, $h = 15$, $n_{a+h} = 39$, $n_a = 83$, $n_{a-h} = 80$, takže modus

$$\hat{x} = 98{,}46.$$

V dokonale souměrném rozdělení četností spadá průměr $\bar{x}$, medián $\tilde{x}$ i modus $\hat{x}$ do jedné hodnoty, která je středem souměrnosti.

V nesouměrném rozdělení četností se tyto tři charakteristiky od sebe liší a je-li rozdělení mírně nesouměrné (obr. 3 nebo 6), ukazuje zkušenost, že medián leží přibližně ve třetině vzdálenosti od průměru k modu, čili osvědčuje se s překvapující přiléhavostí přibližný vztah

$$\bar{x} - \hat{x} = 3(\bar{x} - \tilde{x}), \qquad (29)$$

z něhož je také možno přibližně modus určit, známe-li již průměr a medián. Zvláště u nesouměrných rozdělení četností mívá modus velkou důležitost jako hodnota znaku, která se

v souboru nejčastěji vyskytuje, takže se někdy nazývá typická hodnota.

V některých odvětvích praktické statistiky mají zvláštní oprávnění ještě jiné charakteristiky polohy.

d) Geometrický průměr je r-tá odmocnina ze součinu všech r pozorovaných hodnot znaku vyšetřovaného souboru. Při skupinovém rozdělení četností, kde je n_i hodnot proměnné zastoupeno třídním znakem x_i, bude při l třídách definován geometrický průměr g rovnicí

$$g = (x_1^{n_1} \cdot x_2^{n_2} \dots x_l^{n_l})^{\frac{1}{r}}.$$

Logaritmováním dostaneme

$$\log g = \frac{1}{r} \sum_{i=1}^{l} n_i \log x_i,$$

z čehož patrno, že logaritmus geometrického průměru je aritmetickým průměrem logaritmů jednotlivých hodnot znaku, takže jeho výpočet se tím převádí na metody zavedené již pro aritmetický průměr. Spočívá na všech hodnotách znaku jako aritmetický průměr, ale je na krajové hodnoty méně citlivý. Častého užití se dostalo geometrickému průměru ve statistice cenové a při konstrukci čísel indexních. Pro danou řadu čísel je vždy menší než její aritmetický průměr. Jednoduchý důkaz pro dvě čísla x_1, x_2 od sebe různá vyplývá takto $(\sqrt{x_1} - \sqrt{x_2})^2 > 0$ tedy

$$x_1 + x_2 - 2\sqrt{x_1 x_2} > 0 \text{ a tudíž } \tfrac{1}{2}(x_1 + x_2) > \sqrt{x_1 x_2}.$$

e) Podáme ještě definici harmonického průměru, jehož používá statistická prakse poměrně zřídka. Převratná hodnota $\frac{1}{\gamma}$ harmonického průměru je aritmetickým průměrem převratných hodnot znaku

$$\frac{1}{\gamma} = \frac{1}{r} \sum_{i=1}^{l} \frac{1}{x_i} n_i.$$

Úvahy o harmonickém průměru lze tedy vhodně převésti na úvahy o aritmetickém průměru.

2. Charakteristiky rozptylu. a) Abychom vystihli způsob rozdělení prvků v mezích celkového variačního rozpětí, užíváme nejčastěji směrodatné odchylky $\sigma_x = \sqrt{\mu_{x,2}}$. Čím je při téže jednotce měření hodnot znaku σ_x menší, tím jsou hodnoty proměnné a tedy prvky souboru těsněji seskupeny kolem aritmetického průměru.

b) Pro srovnávání někdy užíváme za účelem eliminování vlivu jednotky měření poměru směrodatné odchylky k průměru, tedy míry relativního rozptylu vyjadřované v procentech. Nazývá se koeficient variační a je dán výrazem

$$v = 100 \frac{\sigma_x}{\bar{x}}. \tag{30}$$

c) Utvoříme-li průměr ze všech odchylek hodnot pozorovaného znaku od některé charakteristiky polohy, nepřihlížejíce při tom ke znaménku odchylek, dostáváme t. zv. průměrnou odchylku ϑ. Přirozeným východiskem je tu medián, takže je dána rovnicí

$$\vartheta = \frac{1}{r} \sum_{i=1}^{l} | x_i - \tilde{x} | n_i. \tag{31}$$

Lze dokázati, že průměrná odchylka je nejmenší, měří-li se odchylky od mediánu [4]. Podle empirického pravidla se obyčejně průměrná odchylka velmi blíží $\frac{4}{5}$ směrodatné odchylky pro souměrné nebo jen mírně nesouměrné rozdělení četností.

d) Také vzdálenost obou kvartilů může sloužiti za míru rozptylu. Rozdíl mezi mediánem a dolním kvartilem $\tilde{x} - \tilde{x}_1$, který se při souměrném rozdělení četností rovná rozdílu mezi horním kvartilem a mediánem $\tilde{x}_2 - \tilde{x}$, dává zřejmě možnost posouditi soustřeďování prvků kolem mediánu. Poněvadž pozorovaná rozdělení četnosti nejsou přesně souměrná, volí se za míru rozptylu poloviční součet obou

uvedených hodnot, který se rovná rozdílu mezi kvartily

$$\tfrac{1}{2}(\tilde{x} - \tilde{x}_1 + \tilde{x}_2 - \tilde{x}) = \tfrac{1}{2}(\tilde{x}_2 - \tilde{x}_1) = \varrho \qquad (32)$$

a nazývá se **kvartilová odchylka**.

e) Výpočet směrodatné odchylky nebo střední odchylky spočívá na odchylkách hodnot znaku od průměru nebo od mediánu. Tím se zavádí do počtu prvek, který není nevyhnutelně nutný, neboť rozptyl členů řady statistické je možno měřiti také jejich absolutními diferencemi bez prostředkování jiné charakteristiky. Tak vznikla míra zvaná **střední diference** Δ.

Předpokládejme, že jsme ze všech hodnot znaku pozorovaných na prvcích souboru s rozsahem r vytvořili řadu neklesající

$$x_1 \leqq x_2 \leqq \ldots \leqq x_r.$$

Utvoříme pak diference

$x_r - x_1,\ x_r - x_2,\ x_r - x_3,\ \ldots,\ x_r - x_{r-3},\ x_r - x_{r-2},\ x_{r-1} - x_{r-1}$

$x_{r-1} - x_1,\ x_{r-1} - x_2,\ x_{r-1} - x_3,\ \ldots,\ x_{r-1} - x_{r-3},\ x_{r-1} - x_{r-2},$

. . . . . .

$x_4 - x_1,\ x_4 - x_2,\ x_4 - x_3,$

$x_3 - x_1,\ x_3 - x_2$

$x_2 - x_1.$

Všechny tyto diference jsou kladné nebo nula a jejich počet je zřejmě

$$(r-1) + (r-2) + \ldots + 3 + 2 + 1 = \frac{r(r-1)}{2}.$$

Sečteme-li nyní všechny diference, zruší se část sčítanců, jak se snadno přesvědčíme, a dostaneme celkový součet S

$$S = (r-1)(x_r - x_1) + (r-3)(x_{r-1} - x_2) + \\ + (r-5)(x_{r-2} - x_3) + \ldots$$

Poslední člen pro sudé $r = 2k$ bude $1 . (x_{k+1} - x_k)$ a pro liché $r = 2k + 1$ je $2 . (x_{k+2} - x_k)$.

Střední diference tedy bude

$$\Delta = \frac{S}{\frac{1}{2}r(r-1)} = \frac{2}{r(r-1)} \sum_{i=1}^{k} (r+1-2i)(x_{r-i+1} - x_i). \quad (33)$$

Výpočet součtu se provede sestavením tabulky

$$\begin{aligned} r-1, &\quad x_r - x_1, \\ r-3, &\quad x_{r-1} - x_2, \\ r-5, &\quad x_{r-2} - x_3, \\ &\ldots\ldots\ldots\ldots \end{aligned}$$

která se k-tým řádkem ukončí, kde k se určí buď ze vztahu $r = 2k$, nebo $r = 2k + 1$. Součinitel $r + 1 - 2i$ pak musí míti hodnotu buď 1 nebo 2.

f) Nejprostším odhadem rozptylu rozdělení četností je variační rozpětí; obsah jeho informace je však poměrně malý.

3. Charakteristiky šikmosti čili kososti. a) O momentové charakteristice šikmosti $\alpha_{x,3}$ jsme se již dostatečně zmínili (str. 23).

b) Míru šikmosti čili nesouměrnosti, nezávislou na jednotce, v níž je měřen pozorovaný znak, můžeme také sestrojiti, stanovíme-li poměr mezi rozdílem odchylek kvartilů od mediánu a vzdáleností obou kvartilů

$$\frac{(\tilde{x}_2 - \tilde{x}) - (\tilde{x} - \tilde{x}_1)}{\tilde{x}_2 - \tilde{x}_1} = \frac{\tilde{x}_1 + \tilde{x}_2 - 2\tilde{x}}{\tilde{x}_2 - \tilde{x}_1} = \tau,$$

při čemž $-1 \leqq \tau \leqq +1$.

Tato míra šikmosti charakterisuje spíše tvar rozdělení četností mezi oběma kvartily a nepřihlíží náležitě k významu hodnot znaku, ležících vně. Může se tudíž v některém případě podstatně lišiti od momentové charakteristiky $\alpha_{x,3}$. Tím se také vysvětluje v případě zvoleného rozdělení četností (41), že hodnota $\tau = -0{,}02$ nasvědčuje souměrnému rozdělení mezi oběma kvartily, kdežto momentová charakteristika udává malou kladnou šikmost.

c) Místo této míry zavedl Pearson výraz $\tau_1 = \frac{\bar{x} - \hat{x}}{\sigma_x}$.

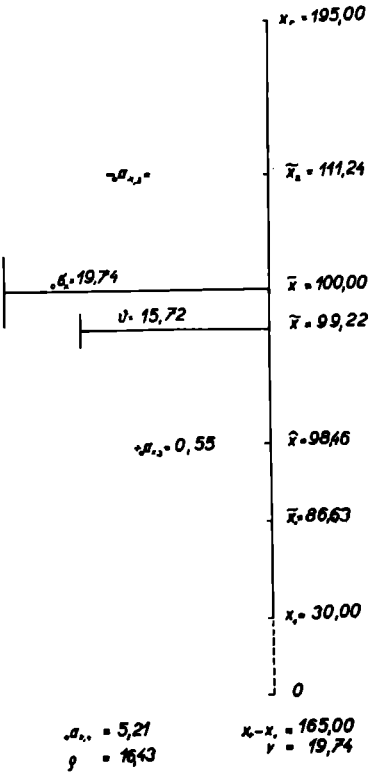
V případě souměrného rozdělení četnosti je $\bar{x} = \hat{x}$ a tudíž $\tau_1 = 0$.

Poněvadž se modus elementárním způsobem nesnadno zjišťuje, nahražuje se někdy čitatel zlomku přibližným $3(\bar{x} - \tilde{x})$ podle rovnice (29), takže touto obměnou dostáváme

$$\tau_2 = \frac{3(\bar{x} - \tilde{x})}{\sigma_x}.$$

Jako je účelno zachovávati jistý pořádek, který jsme zavedli při výpočtu charakteristik a zapisování výsledků, tak se jeví také přehledným stálé schema, na které si zvykneme pro charakterisování pozorovaného souboru. Pro náš numerický příklad je sestavíme a vyplníme.

Do tohoto schematu ovšem zapíšeme jen ty charakteristiky, které jsme potřebovali a tedy počítali. Neznáme-li při skupinovém rozdělení četností největší pozorovanou hodnotu x_r, zapíšeme si třídní znak poslední třídy x_l; x_1 značí buď nejmenší pozorovanou hodnotu znaku, nebo třídní



Obr. 8. Přehledné schema charakteristik.

znak první třídy. $\vartheta \doteq \frac{1}{5}\sigma_x$. Šikmost zapíšeme nahoru nebo dolů, podle toho, je-li záporná či kladná.

(3,15) Tři druhy řad. 1. Rozdělení četností se také nazývá statistickou řadou věcnou, která podává roztřídění pozorovaného souboru podle hodnot nějakého znaku, bez ohledu na čas nebo prostor. Vedle těchto řad se rozeznávají

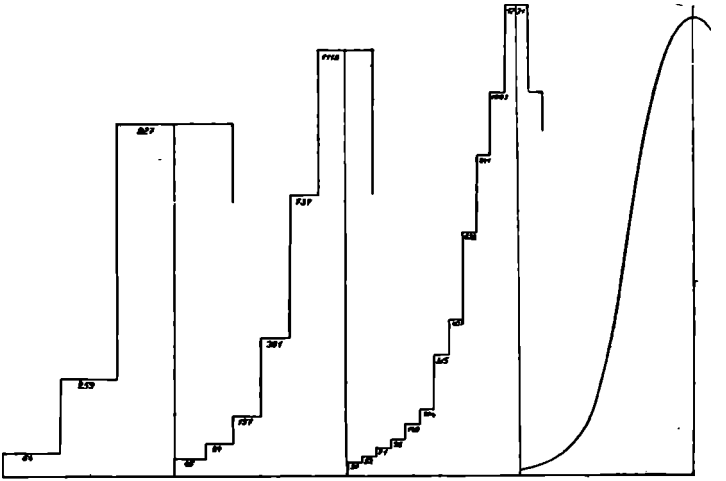
2. řady časové, v nichž jsou jednotlivé hodnoty nebo četnosti uspořádány podle souslednosti časové. Mohou se také nazývati chronologické nebo historické. Časová stupnice, podle níž je řada uspořádána, je dána jednotkou, nejčastěji rok, nebo měsíc, týden, den. (Na př. hodnota dovozu nebo vývozu za každý měsíc.) Znázorňují se chronologickým diagramem.

3. řady místní, kde číselné údaje o jevu pozorovaném v určitém okamžiku jsou uspořádány podle místní příslušnosti (do obce, okresu, země, ...). Znázorňují se obyčejně kartogramem.

(3,16) Od skupinového rozdělení četností ke spojité křivce. Viděli jsme, že při volbě délky intervalu třídního máme značnou volnost. Tvar rozdělení četností pak do jisté míry závisí na této volbě. Budeme sledovati třídění spojitého znaku na pozorovaném souboru dosti velikého rozsahu $r = 10\,000$, který má při délce intervalu $h = 2$ rozdělení četnosti ve druhém sloupci tabulky 3. Provedeme-li přeřazení do tříd dvojnásobné délky intervalu a čtyřnásobné délky intervalu, dostáváme sloupce (3) a (4) tabulky 3.

Roztříděním prvků souboru a získáním rozdělení četností jsme zjistili jak jsou rozděleny prvky vzhledem k vyšetřovanému znaku v daném množství 10 000. Máme v tom však také odpověď na otázku: Vezmeme-li náhodně 10 000 předmětů druhu definovaného statistickou jednotkou a rozdělíme je do skupin, jak často vezmeme prvek do každé z těchto skupin? — Znázorníme-li si hrubé rozdělení (sloupce 4)

histogramem (obr. 9), který vystupuje po stupních nahoru a pak zase sestupuje (což již pro úsporu místa a přehlednost není vykresleno), vidíme, že počet prvků v třídách blízkých průměru je největší a k oběma krajům klesá. Vzhledem k tomu očekáváme, že rozdělíme-li některou ze tříd na dvě,



Obr. 9. Vznik hladšího rozdělení četnosti zkracováním třídního intervalu.

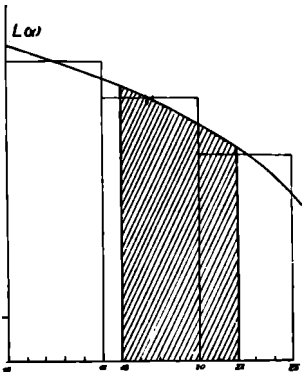
bude část bližší průměru obsahovati více prvků než část vzdálenější. Tak rozdělíme-li interval 162, 164, 166, 168 na dva 162, 164 a 166, 168, vidíme, že jeho četnost 1036 se rozpadne na četnost intervalu bližšího průměru 722 a na četnost 314 intervalu 162, 164 vzdálenějšího od průměru. Podobně je tomu v případech ostatních intervalů. Stupně jsou užší a jejich počet vzrostl, jak je patrno ze zobrazení levé části rozdělení četností histogramem, kde je tedy četnost znázorněna plochou příslušného pravoúhelníka. Opaku-

Tabulka 3.

Výška v cm	Četnost pro interval		
	h	$2h$	$4h$
(1)	(2)	(3)	(4)
154—	37		
156—	52	89	
158—	71		
160—	96	167	256
162—	140		
164—	174	314	
166—	315		
168—	407	722	1036
170—	632		
172—	841	1473	
174—	1003		
176—	1232	2235	3708
178—	1232		
180—	1003	2235	
182—	841		
184—	632	1473	3708
186—	407		
188—	315	722	
190—	174		
192—	140	314	1036
194—	96		
196—	71	167	
198—	52		
200—	37	89	256

jeme-li tento postup dále (sloupec 2), stupně se zúžují a průběh je hladší. Kdyby byl dostatečně velký rozsah souboru a hodnoty znaku pozorovány na dosti velký počet desetinných míst, zmizely by stupně a při nekonečně velkém rozsahu souboru dostali bychom spojitou křivku.

Je tedy jasno, že stupně jsou něčím umělým, neboť vznikají tím, že musíme volit do značné míry libovolně hranice tříd vzhledem k tomu, že znak je měřen různými měrami ať délky nebo váhy či věku, atd. Poněvadž tvoření tříd je libovolné, býváme nuceni rozdělení četností nahradit něčím, co nesouvisí s uspořádáním podléhajícím této libovůli. Můžeme proložit spojitou křivku vrcholky polygonu četnosti nebo jí nahradit také histogram. Spojitá křivka je nezávislá na třídách a proto je obecnější povahy než hrubý polygon. Mimochodem se zmíníme o možnosti užití spojité křivky četnosti, známe-li jen tabulku skupinového rozdělení četností v určitých třídních intervalech a potřebujeme je znáti v třídách utvořených jinak. Dostaneme na př. při sčítání lidu tabulku četností $L(x)$ jen pro pětileté nebo desetileté věkové třídy a potřebujeme k určitému účelu znáti počet osob věku 16—22 let. Z původního materiálu to již není možné, nebo by to bylo při rozsáhlosti souboru příliš nákladné. Úlohu rozřešíme potom tak, že histogram pro pětileté třídní intervaly nahradíme přiléhající křivkou, která uzavírá s osou x plochu stejnou jako s ní uzavírá obrys histogramu (obraz 10) a změříme plochu odpovídající uvedenému intervalu. Druhý je případ užití křivky četností je-li rozsah



Obr. 10. Stanovení četnosti pro změněný interval.

souboru malý, tedy pozorované četnosti třídní malé a vyznačující se nepravidelností. Proložíme tedy křivku, abychom odstranili nahodilé výkyvy a dostali celkový průběh, který by se přibližoval průběhu spojitému, jejž bychom dostali, kdyby rozsah souboru rostl nade všechny meze. Při prokládání křivky pozorovanými hodnotami volnou rukou je třeba velké opatrnosti, neboť může dáti někdy velmi nesprávný odhad ideálního výsledku. Proto se k tomu cíli užívá zvláštních metod matematických.

(4,1) Vznik hlavních typů rozdělení četností.

Typy křivek, které se uplatňují ve statistice, vyvozujeme dvojí cestou: jednak dedukcí pomocí kombinatorických úvah, zabýváme-li se problémy ryzí náhody, jako házení mincí, kostek, ... jednak indukcí, studujeme-li tvary křivek, které se obecně vyskytují při zkoumání souborů velkých rozsahů z různých oborů statistiky.

Obě cesty se doplňují a pomocí modelů sestrojených úvahami kombinatoriky [10], [11] osvětlujeme výsledky pozorování, u nichž můžeme souditi na analogické podmínky vzniku dotyčného jevu (poměr počtu narozených chlapců a děvčat), pro který však opakované provedení nějakého statistického experimentu je vyloučeno.

Uděláme si nejprve představu o spojení mezi náhodným jevem a spojitými křivkami.

Zvolíme si za pokus házení mincí a pozorovaný znak bude počet rubů a líců, které se objeví. Užijeme pracovní hypotésy, že každá mince, kterou házíme, je správná, čili vykazuje příslušnou geometrickou a mechanickou symetrii. Není tudíž důvodu k tomu, aby se rub objevoval u téže mince častěji než líc.

Zaznamenejme si všechny případy, které mohou nastati, když házíme třemi stejnými mincemi. Sestavíme si je podle počtu rubů; označíme-li písmenou „L" líc, „R" rub, pak vidíme:

R R R ..... 1 případ: tři ruby

R R L .....
R L R ..... } 3 případy: dva ruby
L R R .....

R L L .....
L R L ..... } 3 případy: jeden rub
L L R .....

L L L ..... 1 případ: žádný rub

Můžeme tedy sestaviti tabulku:

počet rubů	0	1	2	3	celkem
počet případů	1	3	3	1	8

Kdybychom takto postupovali dále pro čtyři mince, dostali bychom počet případů, čili četnost, vyjádřenou řadou čísel 1, 4, 6, 4, 1 a obecně jak známo řadou binomickou $(1 + 1)^l$.

Tato čísla tvoří pro $l = 0, 1, 2, \ldots$ známý Pascalův trojúhelník [10, 11]. Když bychom prováděli skutečné pokusy, dostaneme ovšem vždy něco jiného. Tak na př. pro 14 mincí dostáváme naší úvahou řadu četností a) 1, 14, 91, 364, 1001, 2002, 3003, 3432, 3003, 2002, 1001, 364, 91, 14, 1 a v dříve již uvedeném případě, kde jsme vykonali 201 vrhů se 14 mincemi, dostáváme tabulku:

b)

počet rubů...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	celkem
počet případů	0	0	1	3	17	23	35	49	35	20	9	8	—	1	0	201

Porovnání můžeme ovšem provésti pomocí relativních četností, čili převedením souboru na rozsah jednotkový. V našem případě pro $l = 14$ je součet absolutních četností $(1 + 1)^{14} = 2^{14} = 16\,384$, takže tímto číslem dělíme každý člen řady a); četnosti pozorované a sestavené právě v tabulce pak dělíme celkovým rozsahem 201.

Dostáváme tak v procentech tyto dvě řady relativních četností

	0	1	2	3	4	5	6	7
a)	0,0001	,0008	,0056	,0222	,0611	,1222	,1833	,2094
	8	9	10	11	12	13	14	
	,1833	,1222	,0611	,0222	,0056	,0008	,0001	

b)	0	1	2	3	4	5	6	7
	0	0	0,0050	,0149	,0846	,1144	,1741	,2438
	8	9	10	11	12	13	14	
	,1741	,0995	,0448	,0398	0	,0050	0	

řada b) se v jistých mezích mění, takže bychom při druhém pokusu o témž počtu vrhů dostali četnosti odlišné.

O mezích těchto odchylek budeme uvažovati později (str. 110).

Úloha:

Jest stanoviti aritmetický průměr a směrodatnou odchylku rozdělení relativních četností daného binomickou řadou

$$(1+1)^l \frac{1}{2^l}:$$

hodnoty znaku: $0 \quad 1 \quad 2 \ldots\ldots\ldots l;$

relat. četnosti: $1\,\frac{1}{2^l} \quad \binom{l}{1}\frac{1}{2^l} \quad \binom{l}{2}\frac{1}{2^l} \cdots \binom{l}{l}\frac{1}{2^l}.$

Násobíme-li hodnoty znaku příslušnými relativními četnostmi a sečteme, dostáváme po vytknutí l

$$l\left\{1 + \binom{l-1}{1} + \binom{l-1}{2} + \ldots + \binom{l-1}{l-1}\right\}\frac{1}{2^l} =$$

$$= l\{1+1\}^{l-1}\frac{1}{2^l} = l\,\frac{1}{2},$$

tedy

$$\bar{x} = \frac{l}{2}.$$

Rozptyl stanovíme, vynásobíme-li čtverce hodnot znaku příslušnými relativními četnostmi a od součtu odečteme $\bar{x}^2$.

$$\left\{1 \cdot \binom{l}{1} + 4\binom{l}{2} + 9\binom{l}{3} + \ldots + l^2\binom{l}{l}\right\}\frac{1}{2^l} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{1 + 2\binom{l-1}{1} + 3\binom{l-1}{2} + 4\binom{l-1}{3} + \ldots + l\binom{l-1}{l-1}\right\}\frac{1}{2^{l-1}} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{1 + \binom{l-1}{1} + \binom{l-1}{2} + \binom{l-1}{3} + \ldots + \binom{l-1}{l-1} + \right.$$

$$\left. + \binom{l-1}{1} + 2\binom{l-1}{2} + 3\binom{l-1}{3} + \ldots + (l-1)\binom{l-1}{l-1}\right\} \times$$

$$\times \frac{1}{2^{l-1}} - \left(\frac{l}{2}\right)^2 =$$

$$= \frac{l}{2}\left\{\left(\frac{1}{2} + \frac{1}{2}\right)^{l-1} + \frac{1}{2}(l-1)\left(\frac{1}{2} + \frac{1}{2}\right)^{l-2}\right\} - \left(\frac{l}{2}\right)^2 =$$

$$\frac{l}{2}\left\{1 + \frac{l-1}{2}\right\} - \left(\frac{l}{2}\right)^2 = \frac{l}{2} - \frac{l}{4} = \frac{l}{4}.$$

Je tudíž

$$\sigma_{x^2} = \tfrac{1}{4}l$$

Zabývejme se nyní blíže rozdělením četností v případě l mincí a zkoumejme, jaké dostaneme rozdělení, jestliže l se stále zvětšuje.

Relativní četnosti v případě obecném l mincí jsou tedy vyjádřeny jednotlivými členy řady

$$\frac{1}{2^l}\left\{1 + \binom{l}{1} + \binom{l}{2} + \ldots + \binom{l}{l}\right\},$$

takže četnost x rubů čili příznivých výsledků, je

$$\frac{1}{2^l}\binom{l}{x} = \frac{1}{2^l}\frac{l!}{x!\,(l-x)!}.$$

Četnost $x + 1$ rubů je dána následujícím členem, a dostáváme ji z předchozího výrazu, násobíme-li jej zlomkem $\frac{l-x}{x+1}$. Pokud bude $l - x > x + 1$ čili $x < \frac{l-1}{2}$, bude následující četnost větší než předcházející. Učiníme pro zjednodušení další úvahy předpoklad, $l = 2\nu$. Při sudém l je nejčetnějším případ ν příznivých výsledků (viděli jsme na příklad, že pro $l = 4$, je

největší četnost pro 2 ruby); jeho relativní četnost je dána výrazem

$$y_0 = \frac{1}{2^{2\nu}} \frac{(2\nu)!}{\nu!\,\nu!}.$$

To by byla v grafickém znázornění největší pořadnice, od níž se svažuje mnohoúhelník relativních četností na obě strany souměrně. Vezměme tedy v úvahu relativní četnost, která přísluší $\nu + x$ příznivým výsledkům, která je

$$y_x = \frac{1}{2^{2\nu}} \frac{(2\nu)!}{(\nu + x)!\,(\nu - x)!}$$

a utvořme podíl

$$\frac{y_x}{y_0} = \frac{\nu\,(\nu - 1) \ldots (\nu - x + 1)}{(\nu + 1)\,(\nu + 2) \ldots (\nu + x)},$$

který můžeme dělením čitatele i jmenovatele ν^x uvésti na tvar

$$\frac{y_x}{y_0} = \frac{\left(1 - \frac{1}{\nu}\right)\left(1 - \frac{2}{\nu}\right) \cdots \left(1 - \frac{x-1}{\nu}\right)}{\left(1 + \frac{1}{\nu}\right)\left(1 + \frac{2}{\nu}\right) \cdots \left(1 + \frac{x-1}{\nu}\right)\left(1 + \frac{x}{\nu}\right)} \qquad (34)$$

Stanovíme přibližnou hodnotu tohoto zlomku za předpokladu, že ν je veliké u porovnání s x a to tak, že můžeme zanedbati $\left(\frac{x}{\nu}\right)^2$ u srovnání s $\frac{x}{\nu}$. Poněvadž nemusíme vzhledem k pravidlu šesti sigma přihlížeti na jedné straně symetrického rozdělení k hodnotám $x > 3\sigma_x$, může býti náš předpoklad splněn. Při našem binomickém rozdělení je $\sigma_x = \sqrt{\frac{\nu}{2}}$ a tedy $\frac{x}{\nu}$ je pak $\frac{3}{\sqrt{2\nu}}$, což je při velkém ν číslo malé. Můžeme nyní použíti, za uvedeného předpokladu, rozvojů jednotlivých činitelů v čitateli i ve jmenovateli (34) v logaritmické řady (Čech sv. 20, str. 91) podle známého vztahu

$$\lg (1 + \varepsilon) = \varepsilon - \tfrac{1}{2}\varepsilon^2 + \tfrac{1}{3}\varepsilon^3 - \tfrac{1}{4}\varepsilon^4 + \ldots$$

a podržíme vždy jen první člen.

Tak dostaneme přibližné vyjádření pro logaritmus zlomku (34)

$$\lg \frac{y_x}{y_0} = -\frac{2}{\nu}\,[1 + 2 + 3 + \ldots + (x - 1)] - \frac{x}{\nu} =$$

$$= -\frac{x\,(x-1)}{\nu} - \frac{x}{\nu} = -\frac{x^2}{\nu}$$

a přejdeme-li od logaritmu k číslu

$$y_x = y_0 e^{-\frac{x^2}{\nu}}.$$

Vzhledem k tomu, že $\nu = 2\sigma_x^2$, můžeme konečně psáti

$$y_x = y_0 e^{-\frac{x^2}{2\sigma_x^2}}. \qquad (35)$$

což je výraz pro t. zv. normální funkci Laplace-Gaussovu. Pořadnice y_0, která odpovídá hodnotě $x = 0$, je maximální pořadnicí, což je zřejmo také z toho, že při $x = 0$ je

$$e^{-\frac{x^2}{2\sigma_x^2}} = \frac{1}{e^{\frac{x^2}{2\sigma_x^2}}} = 1$$

a při jakékoliv jiné hodnotě x bude

$$\frac{1}{e^{\frac{x^2}{2\sigma_x^2}}} < 1.$$

Tak jsme dospěli postupným zvětšováním počtu hodnot znaku až ke znaku spojitému a od rozpojitého rozdělení četností ke spojitému, vyjádřenému symetrickou křivkou normální.

Jsou však také jiné křivky, vyjadřující rozdělení četností jevů, které nejsou symetrické; můžeme je odvoditi podobnými úvahami. Tak nám dávají známé úvahy kombinatoriky počet případů, v nichž se objeví při házení osmi kostkami jednotky nebo dvojky binomickým rozvojem: $(1 + 2)^8$, který nám dává tato čísla:

počet jednotek nebo dvojek..	0	1	2	3	4	5	6	7	8	celkem
četnost.....	256	1024	1792	1792	1120	448	112	16	1	3^8

Asymetrie je patrna; při množství pokusů bychom na př. dostali v průměru jen jednou ze $6561 = 3^8$ vrhů případ, že by všechny kostky dávaly na horní straně buď jednotky nebo dvojky.

ČÁST III.

(5,1) Teorie náhodného výběru. (Znak alternativní.) Hodnota relativní četnosti v základním souboru — pravděpodobnost.

Jakmile přecházíme od popisných úkolů k bližšímu vysvětlování pozorovaných jevů hromadných, opíráme se o pojem pravděpodobnosti a věty odvozené počtem pravděpodobnosti. Při t. zv. statistické definici pravděpodobnosti vycházíme od posloupnosti jevů. Procházíme-li zápisy v matrice nějakého většího města, které jsou vedeny časově za sebou třeba po dvacet let a zaznamenáváme porody podle znaku pohlaví, takže označujeme chlapce *c*, děvčata *d*, dostaneme posloupnost, jejíž členy opatříme pořadovými čísly (v druhém řádku)

c	*d*	*c*	*d*	*d*	*c*	*c*	*c*	*d*	*c*	*d*	*d*	*d*	*c*	*c*	*d*	*c*	...
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...
1	1	2	2	2	3	4	5	5	6	6	6	6	7	8	8	9	...

Abychom nabyli přehledu o četnosti narozených chlapců v určitém úseku posloupnosti, sčítáme v něm písmena *c*. Jedná-li se o úsek od začátku až do některého pořadového čísla i, napíšeme pod něj v třetím řádku zjištěný počet písmen *c*, který označujeme jako absolutní četnost n_i, takže v naší posloupnosti jsou četnosti chlapců postupně

$n_1 = 1, n_2 = 1, n_3 = 2, \ldots, n_7 = 4, n_8 = 5, \ldots, n_{16} = 8, n_{17} = 9, \ldots$

Vidíme, že k původní posloupnosti porodů náleží posloupnost absolutních četností znaku *c* a tudíž také posloupnost relativních četností $f_i = \frac{n_i}{i}$, která má pro naše další účele zvláštní význam, neboť jsme viděli, že základní formou podávání výsledků statistického šetření je relativní četnost

znaku v pozorovaném souboru. Je zřejmo, že absolutní četnosti jsou mezi 0 a i, takže platí nerovnosti $0 \leqq n_i \leqq i$ a pro relativní četnosti tedy $0 \leqq f_i \leqq 1$.

Posloupnost relativních četností je v našem případě

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{3}, \frac{2}{4}, \frac{2}{5}, \frac{3}{6}, \frac{4}{7}, \frac{5}{8}, \frac{5}{9}, \frac{6}{10}, \frac{6}{11}, \cdots$$

Úsek posloupnosti, který jsme uvedli, je zcela nepatrný. Kdybychom sledovali v dlouhém úseku pěti let, šesti, sedmi, osmi, ... let vývoj čísel f_i, pozorovali bychom, že se stále blíží určitému číslu na př. 0,51, od něhož se liší na desetinných místech vždy vzdálenějších.

Číslo 0,51 dostáváme pro celý soubor, který představuje posloupnost prvků za celých dvacet let. Tento soubor je vyššího řádu než soubory částečné, jež tvoří posloupnosti pozorované za kratší časové úseky. Považujeme jej za soubor základní.

Relativní četnost v základním souboru nazýváme statistickou pravděpodobností, kterou budeme označovati p.

Základní soubor si budeme představovati jako soubor, jehož prvky jsou dobře promíchány, což znamená, že ve všech jeho částech je relativní četnost pozorovaného znaku přibližně táž. Budeme nejprve předpokládati, že známe relativní četnost znaku c v základním souboru čili pravděpodobnost p. Rozsah základního souboru je N.

Budeme bráti náhodně ze základního souboru výběry o r prvcích, tak jako bereme kuličky z osudí. Takové soubory budeme nazývat náhodné výběry. Na vyňatém prvku zjistíme, má-li pozorovaný znak, a zase jej vrátíme do základního souboru, takže se v něm p nemění. V náhodných výběrech rozsahu r prvků se bude vyskytovat různý počet prvků s pozorovaným znakem c, který označíme x. Budou výběry, v nichž nebude ani jeden prvek se znakem c, tedy $x = 0$, v jiných bude $x = 1, 2 \ldots$ a v některých $x = r$. Dělíme-li tento počet prvků se znakem c rozsahem výběru r, dostaneme relativní četnost $\frac{x}{r}$. Všech možných výběrů

a tedy také hodnot x bude $\binom{N}{r}$, neboť tolika způsoby lze kombinovat N prvků po r; představíme si, že tyto hodnoty tvoří nový soubor. Chceme především stanovit, jaké jsou v něm relativní četnosti jednotlivých hodnot $x = 0, 1, 2, \ldots$ $\ldots, r$, tedy konkretních kombinací s x prvky znaku c.

Z počtu pravděpodobnosti známe pravděpodobnost, že nastane v souboru r pokusů právě x-krát jev, jehož pravděpodobnost je p. Je dána t. zv. Newtonovou formulí

$$P_r(x) = \binom{r}{x} p^x q^{r-x}, \tag{36}$$

kde $q = 1 - p$.

Bude tedy celé rozdělení četností určeno všemi členy $P_r(x)$, t. j. pro $x = 0, 1, 2, \ldots, r$, což jsou jak známo členy binomického rozvoje

$$(q + p)^r = q^r + rpq^{r-1} + \binom{r}{2} p^2 q^{r-2} + \ldots + \\ + \binom{r}{x} p^x q^{r-x} + \ldots + p^r, \tag{37}$$

jak se odvozuje v počtu pravděpodobnosti [10], [11] pro pravděpodobnosti opakovaných jevů. Jednotlivé členy mají charakter statistických pravděpodobností, jak je patrno z odvození, neboť jsme je nedostali jako výsledky skutečně provedených výběrů.

(5,2) Binomické rozdělení četností, jeho průměr a rozptyl. Rozdělení četností, jehož třídní četnosti jsou úměrné členům tohoto rozvoje, se také nazývá rozdělení Bernoulliho.

Jeho důležitost není jenom v tom, že udává nejpravděpodobnější rozdělení výběrů z osudí, nýbrž vystihuje typ rozdělení relativních četností, které dostáváme při nejjednodušších operacích náhodného výběru ve statistice. Tak považuje na př. biolog rozvoj (37) za teoretické rozdělení

relativních četností chlapců v náhodných výběrech o rozsahu r porodů. Pojistný technik na př. považuje rozvoj (37) za teoretické rozdělení ročních měr úmrtnosti v náhodných výběrech rozsahu r mužů téhož věku, třeba 25 roků. Při tom nutno zdůrazniti, že tyto výběry jsou brány stále za týchž podmínek, zde ze souboru mužů stále stejného složení vzhledem k znakům, které mohou míti vliv na úmrtnost, tedy vzhledem k povolání, zdravotnímu stavu a pod. Předpoklad stále stejných podmínek čili stálého p je podkladem základním při odvozování Bernoulliova rozdělení; jinými slovy provádění jednoduchého náhodného výběru předpokládá, že základní pravděpodobnost p výskytu znaku zůstává konstantní od výběru k výběru, v němž jednotlivé prvky jsou vzájemně nezávislé, t. j. na zahrnutí prvku do výběru nemá významného vlivu zahrnutí prvku předcházejícího.

Nejpravděpodobnější počet x' prvků, se znakem c ve výběru rozsahu r najdeme, utvoříme-li poměr obecného člena rozvoje, k předcházejícímu a pak k následujícímu; tyto dva poměry budou rovny nebo větší než 1.

$$\frac{r-x+1}{x}\frac{p}{q} \geqq 1 \quad \text{čili} \quad x \leqq pr+p$$

a stejně druhý poměr

$$\frac{x+1}{r-x}\frac{q}{p} \geqq 1 \quad \text{čili} \quad x \geqq pr-q.$$

Z toho plyne, že pro celá čísla x je největší hodnota určena nerovnostmi

$$pr-q \leqq x' \leqq pr+p$$

nebo vzhledem ku $p+q=1$

$$pr+p-1 \leqq x' \leqq pr+p,$$

takže zanedbáme-li pravý zlomek, nejčetnější hodnota počtu prvků se znakem c je pr. Jsou-li $pr-q$ a $pr+p$ čísla celá, existují dva stejné největší členy rozvoje. [Ukažte, že první dva členy rozvoje $(\frac{5}{6}+\frac{1}{6})^5$ jsou stejné.]

Odvodíme si pro toto rozdělení četností první dvě charakteristiky, jimž budeme říkati parametry, ježto jsou to hodnoty v základním souboru, kde relativní četnost znaku pozorovaného je rovna pravděpodobnosti p. Hodnotu průměru v základním souboru označíme $\mathfrak{E}(x)$ a dostaneme podle definice, je-li x hodnota znaku

$$\mathfrak{E}(x) = \sum_{x=0}^{r} \binom{r}{x} p^x q^{r-x} \,.\, x =$$

$$= \sum_{x=0}^{r} \frac{r!}{x!\,(r-x)!} p^x q^{r-x} \,.\, x = \sum_{x=0}^{r} \frac{r!}{(x-1)!\,(r-x)!} p^x q^{r-x} =$$

$$= rp \sum_{x=1}^{r} \frac{(r-1)!}{(x-1)!\,(r-x)!} p^{x-1} q^{r-x} = rp,$$

neboť

$$\sum_{x=1}^{r} \binom{r-1}{x-1} p^{r-1} q^{r-x} = (p+q)^{r-1} = 1.$$

Dostáváme tudíž výsledek

$$\mathfrak{E}(x) = rp. \tag{38}$$

Kdyby hodnoty znaku byly $\frac{x}{r} = f$, je patrno, že bychom dostali průměr $\mathfrak{E}(f) = p$. Abychom odvodili v tomto rozdělení četností hodnotu rozptylu $\sigma^2(x)$, utvoříme součet čtverců odchylek od průměru $\xi = x - rp$, takže podle definice bude

$$\begin{aligned} \sigma^2(x) &= \sum_{x=0}^{r} \binom{r}{x} p^x q^{r-x} (x - rp)^2 = \\ &= \sum_{x=0}^{r} \binom{r}{x} p^x q^{r-x} (x^2 - 2xrp + r^2p^2). \end{aligned} \tag{39}$$

Místo x^2 užijeme identického výrazu $x^2 = x + x\,(x-1)$, takže první člen můžeme psáti

$$\sum_{x=0}^{r} \frac{r!}{x!\,(r-x)!}\, p^x q^{r-x} x + r(r-1)\, p^2 \sum_{x=2}^{r} \frac{(r-2)!}{(x-2)!\,(r-x)!} \times$$
$$\times\, p^{x-2} q^{r-x} = rp + r(r-1)\, p^2,$$

druhý člen je

$$2rp \sum_{x=0}^{r} \binom{r}{x} p^x q^{r-x} \,.\, x = 2r^2 p^2$$

a třetí člen

$$r^2 p^2 \sum_{x=0}^{r} \binom{r}{x} p^x q^{r-x} = r^2 p^2$$

z čehož plyne, že rozptyl

$$\sigma^2(x) = rp + r(r-1)\, p^2 - 2r^2 p^2 + r^2 p^2 =$$
$$= rp - rp^2 = rp\,(1-p) = rpq. \qquad (40)$$

Uvažujeme-li odchylky relativní četnosti znaku, od pravděpodobnosti výskytu znaku $\frac{x}{r} - p$ dostaneme průměr čtverců těchto odchylek, dělíme-li výraz (39) čtvercem r^2, takže příslušný rozptyl je dán zlomkem $\sigma^2(f) = \frac{pq}{r}$. Je tudíž patrno, že rozptyl absolutních četností roste s rostoucím rozsahem r výběru, kdežto rozptyl relativních četností klesá s rostoucím rozsahem r výběru.

Podotkneme ještě, že bychom dostali obdobně rozptyl pro případ, t. zv. hypergeometrického rozdělení četností (str. 90), jež vystihuje braní výběrů ze základního souboru tím způsobem, že se prvky nevrací zpět. Rozptyl pak je dán výrazy

$$rpq\left(1 - \frac{r}{N}\right) \text{resp. } pq\left(\frac{1}{r} - \frac{1}{N}\right).$$

Tento případ přechází v binomický, je-li rozsah základního souboru velmi velký $N \to \infty$. Také tyto výrazy pro rozptyl přecházejí pak na (40) resp. (40′).

Budeme dále zkoumati, zda není možno udati pro odchylky $\frac{x}{r} - p$, tedy odchylky relativních četností znaku c ve výběrech $\frac{x}{r}$ od relativní četnosti v základním souboru p, takové hranice, v nichž bude většina všech možných výsledků. K tomu cíli si napřed odvodíme důležitou větu.

(5,3) Věta Bienaymé-Čebyševova. Představme si, že máme napozorováno množství hodnot statistické proměnné $x_1, x_2, \ldots, x_l$ s relativními četnostmi resp. $\nu_1, \nu_2, \ldots, \nu_l$, takže $\nu_1 + \nu_2 + \ldots + \nu_l = 1$. Je-li jejich průměr $\bar{x}$ a označíme zase odchylky $x_i, -\bar{x} = \xi_i$, bude rozptyl

$$\sigma_x^2 = \nu_1 \xi_1^2 + \nu_2 \xi_2^2 + \ldots + \nu_l \xi_l^2.$$

Rozdělme nyní odchylky ξ_i na takové, které nedosahují numericky určitého násobku směrodatné odchylky $\tau\sigma_x$, při čemž $\tau > 1$ a na ostatní $|\xi_i| \geqq \tau\sigma_x$. Relativní četnost prvních odchylek označíme P_τ, takže relativní četnost ostatních bude $1 - P_\tau$. Můžeme pak psáti rovnici pro rozptyl

$$\sigma_x^2 = \sum_{i=1}^{l'} \xi_i^2 \nu_i + \sum_{j=l'+1}^{l} \xi_j^2 \nu_j.$$

Kde se první součet vztahuje na všechna ξ_i, která nedosahují $\tau\sigma_x$ a druhý součet na všechna ξ_j, která se mu rovnají a převyšují. Poněvadž máme všechny sčítance obou součtů kladné nebo rovny nule, je

$$\sigma_x^2 \geqq \sum_{j=l'+1}^{l} \xi_j^2 \nu_j.$$

Platí tudíž zřejmě nerovnost

$$\sigma_x^2 > \sum_{j=l'+1}^{l} \tau^2 \sigma_x^2 \nu_j.$$

Vzhledem k tomu, že jsme označili

$$\sum_{j=l'+1}^{l} \nu_i = 1 - P_\tau$$

je také

$$\sigma_x^2 > \tau^2 \sigma_x^2 (1 - P_\tau) \text{ čili } \frac{1}{\tau^2} > 1 - P_\tau$$

a konečně

$$P_\tau > 1 - \frac{1}{\tau^2}, \tag{41}$$

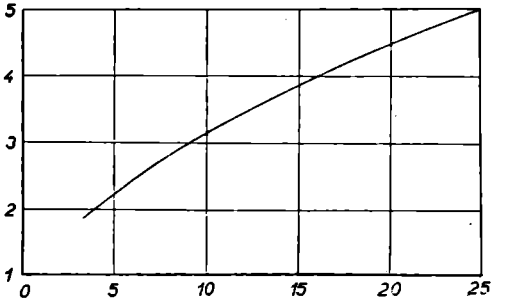
což znamená, že relativní četnost prvků, jejichž hodnota znaku se bude odchylovati od průměru méně než o $\tau\sigma_x$ je větší než $1 - \frac{1}{\tau^2}$. Tato věta se nazývá kriteriem Bienaymé-Čebyševovým.

Všimneme si ještě, že pravědpodobnosti odchylek podle věty Bienaymé-Čebyševovy mají povahu obecnou, která nezávisí nijak na tvaru rozdělení četností. Za to však jsou tyto pravděpodobnosti určeny v úzkých mezích často nepostačujících, neboť je tu udána dolní mez stanovením, že pravděpodobnost odchylky v mezích τ-násobné směrodatné odchylky je větší než $1 - \frac{1}{\tau^2}$. Vzniká zase otázka, jak blízko je tato dolní mez skutečné hodnotě pravděpodobnosti. Tato otázka má praktický význam, neboť je-li tato mez značně níže než skutečná pravděpodobnost, musíme provésti k dosažení uspokojivého výsledku zbytečně mnohem více pozorování, než kdybychom znali skutečnou pravděpodobnost. Podle věty Bienaymé-Čebyševovy leží víc než $1 - \frac{1}{\tau^2}$ z celkového počtu r prvků souboru v mezích $\bar{x} \pm \tau\sigma_x$ (kde ovšem $\tau \geqq 1$) a tato věta platí pro jakoukoliv množinu konečných čísel, bez ohledu na to, jak byla získána. Pro několik hodnot τ si sestavíme přehled:

τ	1	$1\frac{1}{2}$	2	3	4
$1-\frac{1}{\tau^2}$	0	0,56	0,750	0,889	0,937

Známe-li tedy $\bar{x}$ a σ_x, můžeme hned říci, že víc než 75% čísel leží v intervalu $\bar{x} \pm 2\sigma_x$, čili méně než 25% se liší od $\bar{x}$ o více než $2\sigma_x$ atd.

Také vyplývá z věty B.-Č., že při rozsahu $r = 4$ budou všechny prvky souboru v mezích $\bar{x} \pm 2\sigma_x$, neboť jich tam bude víc než $1 - \frac{1}{4}$, tedy ze čtyř více než tři čtvrtiny.



Obr. 11. Hodnoty τ, pro něž všechny prvky jsou v intervalu $x \pm \tau\sigma_x$.

Podobně pro $r = 10$ vidíme, že budou všechny prvky v intervalu $\bar{x} \pm 3{,}16\sigma_x$, neboť jich tam bude víc než $1 - \frac{1}{10}$ a pod. Můžeme si graficky znázorniti obory, v nichž jsou podle věty B.-Č. obsaženy všechny prvky souboru; budou vyznačeny křivkou $\tau^2 = r$ (obr. 11).

(5,4) Teorém Bernoulliův. Budeme nyní s hlediska kriteria B.-Č. uvažovati zmíněnou již úlohu, která je jedním z uhelných kamenů moderní statistiky, totiž najíti pravděpodobnost, že odchylka relativních četností $\left|\frac{x}{r} - p\right|$ bude menší než libo-

volné kladné číslo ε. Zvolíme tedy $\varepsilon = \tau\,\sigma(f)$, kde $\sigma(f) = \sqrt{\frac{pq}{r}}$, potom platí podle věty Bienaymé-Čebyševovy, že pro

$$|\xi_i| \geqq \tau\,\sigma(f) \text{ bude } P_\tau \leqq \frac{1}{\tau^2} \text{ čili } P_\tau \leqq \frac{pq}{r\varepsilon^2} \text{ neboť } \frac{1}{\tau} = \frac{\sigma(f)}{\varepsilon}.$$

Může tedy býti pravděpodobnost P_τ pro rostoucí r při určitě zvoleném ε libovolně malá. Naopak pro pravděpodobnost $1 - P_\tau$ že $\left|\frac{x}{r} - p\right| < \varepsilon$ bude platit $1 - P_\tau > 1 - \frac{1}{\tau^2}$ čili

$$1 - P_\tau > 1 - \frac{pq}{r\varepsilon^2}. \tag{42}$$

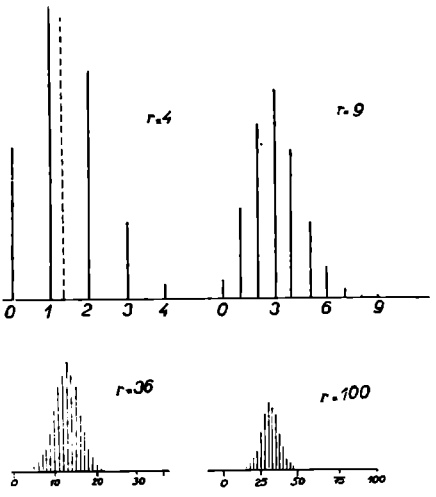
Tato pravděpodobnost se blíží 1, když r roste nad všechny meze. Odhad na pravé straně (42) můžeme provésti nezávisle na určité hodnotě p, neboť součin pq nemůže býti větší než $\frac{1}{4}$, vzhledem k tomu, že $p + q = 1$, takže bude $1 - P_\tau > 1 - \frac{1}{4r\varepsilon^2}$. Tento odhad je přirozeně slabší.

Tak jsme dospěli k teorému Bernoulliovu, který je jedním ze základních pilířů statistiky. Výraz (42) vyjadřuje teorém Bernoulliův jako větu o mezní hodnotě nejjednodušší formou. Osvětluje otázku, jak se blíží relativní četnost znaku ve výběru o r prvcích své hodnotě v základním souboru, t. j. konstantní pravděpodobnosti p, když rozsah r roste a vyslovíme jej takto:

Je-li p pravděpodobnost výskytu znaku pro každý prvek náhodného výběru rozsahu r, pak se pravděpodobnost P_τ odchylky $\frac{x}{r} - p$ relativní četnosti znaku ve výběru od hodnoty p v základním souboru, která se rovná libovolně malému kladnému číslu ε, blíží k nule jakožto limitě, roste-li rozsah náhodného výběru r nade všechny meze. Pravděpodobnost $1 - P_\tau$, že odchylka relativní četnosti znaku ve

výběru $\frac{x}{r}$ od hodnoty p v základním souboru bude menší než ε, se blíží 1 neboli jistotě.

Způsob, jímž jsme přešli od rozdělení četností absolutních x k rozdělení relativních četností $\frac{x}{r} = f$ můžeme považovati za transformaci souřadnic, která sesunuje k sobě úsečky funkce $P_r(x)$ v poměru $r : 1$. Průměr transformovaného rozdělení je konstanta p a rozptyl $\sigma^2(f) = \frac{\sigma^2(x)}{r^2} = \frac{pq}{r}$, neboť rozptyl se mění se čtvercem úseček. Rozptyl tedy klesá k nule s rostoucím r. Rozdělení $P_r(f)$ je znázorněno v obr. (12) pro $p = \frac{1}{3}$; pro $r = 100$ bylo možno zobrazit jeň každou druhou pořadnici. Ubývání rozptylu tu jasně vidíme a současně se jeví, můžeme říci, zhušťování rozdělení četností, čímž je vyjádřena podstata Bernoulliho teorému.



Obr. 12. Zhušťování binomického rozdělení četností a klesající rozptyl.

Můžeme jej také formulovati větou:

Relativní četnost nějakého znaku, zjištěná v náhodném výběru rozsahu r na sobě nezávislých prvků, se blíží hodnotě p v základním souboru až na odchylku (chybu) ε napřed danou s pravděpodobností, která se může zvětšováním rozsahu r přiblížiti libovolně blízko číslu 1.

V tomto smyslu tudíž representuje náhodný výběr rozsahu r celý soubor všech prvků, odpovídajících pojmu určujícímu statistickou jednotku, tím lépe, čím je rozsah výběru r větší. Je to základní věta o větší bezpečnosti delší statistické řady, která tvoří v podstatě obsah t. zv. zákona velkého čísla. Věta o větší bezpečnosti delší statistické řady dává oprávnění principu statistické indukce; mohla by býti označena také jako věta o větší bezpečnosti závěru provedeného statistickou indukcí na základě náhodného výběru o větším rozsahu než na základě výběru o menším rozsahu.

Zákon velkého čísla souvisí přímo s principem stejnotvárnosti přírodního dění, podle něhož za podobných okolností jev probíhá podobně. Předpokládá se tedy, že stejné skupiny (komplexy) příčin mají za následek stejné pochody. Abychom však z teorému matematicky odvozeného mohli činit závěry na skutečné dění, musíme učinit ještě další krok.

Budeme se dovolávat zkušenosti, že v souborech, které mají povahu našeho základního souboru, pozorujeme skutečně zřídka prvky o znaku s malou relativní četností. Vyvodíme z toho potom závěr, že velké odchylky $\frac{x}{r} - p$ se budou u statistických souborů rovněž jen zřídka vyskytovat. To je podstatný obsah věty Cournotovy a jeho formulace zákona velkého čísla.

Na konec nám zbývá určiti celkovou relativní četnost všech těch prvků našeho nového (myšlenkového) souboru rozsahu $\binom{N}{r}$, u nichž se vyskytuje znak c nejméně $(rp - \xi_0)$-krát a nejvýše $(rp + \xi_0)$-krát, neboli u nichž je relativní četnost znaku c v mezích od $p - \frac{\xi_0}{r}$ do $p + \frac{\xi_0}{r}$. Pro první případ dostaneme hledanou celkovou relativní četnost $P_r(\bar{x} - \xi_0, \bar{x} + \xi_0)$, když sečteme všechny relativní četnosti $P_r(x)$ pro hodnoty x v uvedeném intervalu. Pro druhý případ obdobně $P_r(p - z_0\, p + z_0)$ dostaneme sečtením přísluš-

ných hodnot $P_r(f)$; při tom je ovšem $P_r(x) = P_r(f)$, $\frac{\xi_0}{r} = z_0$. Vzhledem k tomu, že výpočet jednotlivých členů je dosti pracný a tedy také jejich součet

$$P_r(\bar{x} - \xi_0, \bar{x} + \xi_0) = \sum_{\bar{x}-\xi_0}^{\bar{x}+\xi_0} P_r(x) = [P_r(\bar{x} + \xi_0) +$$

$$+ P_r(\bar{x} - \xi_0)] + [P_r(\bar{x} + \xi_0 - 1) + P_r(\bar{x} - \xi_0 + 1)] +$$

$$+ \ldots + [P_r(\bar{x} + 1) + P_r(\bar{x} - 1)] + P_r(\bar{x})$$

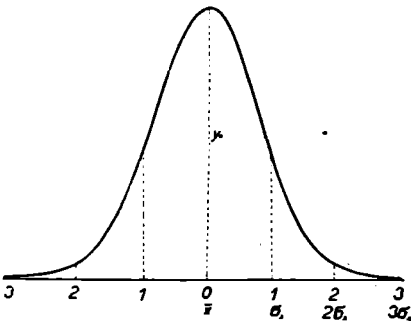
najdeme dále — rovnice (52) nebo (53) — vyhovující řešení přibližné.

Příklad: relativní četnost narozených chlapců v základním souboru je $p = 0{,}513$ čili 51,3%. Průměrná absolutní četnost chlapců narozených ročně v místě, kde se rodí ročně $r = 100$ dětí, je tedy $rp = 51{,}3$; v místě, kde se rodí ročně 10 000 dětí, bude $rp = 5130$. Směrodatná odchylka činí v prvém případě $\sqrt{rpq} = \sqrt{0{,}513 \times 0{,}487r} = 0{,}5\sqrt{r} = 5$, kdežto v druhém případě 50. Směrodatná odchylka relativních četností chlapců však klesá, neboť je v prvém případě $\sqrt{\frac{pq}{r}} = \frac{0{,}5}{10} = 0{,}05$, tedy 5%, kdežto v druhém případě jen 5 promille. Výsledek se považuje za tím přesnější, čím má menší rozptyl a tedy také, čím má menší směrodatnou odchylku. Je z toho zřejmo, že čím jsou náhodné výběry většího rozsahu, tím dávají výsledek bližší hodnotě v základním souboru.

(6,1) Křivky rozdělení četností. (Křivka Laplace-Gaussova.)

Lze říci, že Bernoulli začal studovat binomické rozdělení četností a vyjádřil jednu jeho zvláštní vlastnost ve větě po něm pojmenované, která ukazuje, že vytkneme-li libovolně malý interval kolem hodnoty p a určíme si číslo libovolně blízké jednotce, pak můžeme zvoliti soubor o dosti

velkém počtu prvků, takže relativní četnost pozorovaného znaku padne do zvoleného intervalu s určenou pravděpodobností. Nahraditi binomické rozdělení spojitou křivkou se podařilo Laplaceovi (1812), takže bylo úplně dáno souměrnou zvonovitou křivkou $e^{-\xi^2}$, jejíž pořadnice klesají od průměru tak, že se jejich přirozené logaritmy (se záporným znaménkem) chovají jako čtverce vzdálenosti od průměru (viz obraz 13).



Obr. 13. Křivka Laplace-Gaussova.

Odvodili jsme si již z binomického rozdělení četností spojitou křivku Laplace-Gaussovu, čili normální ve zcela zvláštním případu, kde základní relativní četnosti při alternativním znaku byly sobě rovny, tedy $p = q = \frac{1}{2}$. Lze však ukázati, že obecné rozdělení binomické $(p + q)^r$ se blíží pro velká r křivce normální. Abychom tento postup naznačili, vyjádříme členy binomického rozdělení (37) hodnotami $y(\xi)$ v jednotkových intervalech tak, že pro rozdíl mezi průměrem $\bar{x} = rp$ a četností x znaku ve výběru rozsahu r zvolíme symbol ξ. Potom jednotlivé členy rozdělení četností budou

$$y(\xi) = \frac{r!}{(pr + \xi)!\,(qr - \xi)!}\, p^{pr+\xi} q^{qr-\xi}. \qquad (43)$$

K přibližnému vyjádření faktoriel použijeme Stirlingovy formule

$$n! = n^n e^{-n}\sqrt{2\pi n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \dots\right).$$

Užijeme-li jen prvního členu této řady, dostaneme přibližnou hodnotu, která se rovná přesné hodnotě dělené nějakým číslem mezi 1 a $1 + \frac{1}{10n}$. Stačí tudíž většinou toto přiblížení pro n, která přicházejí v úvahu. S tímto přiblížením pak dostáváme

$$(pr + \xi)! = (pr + \xi)^{pr+\xi} e^{-(pr+\xi)}\sqrt{2\pi (pr + \xi)} =$$
$$= (pr)^{pr+\xi}\left(1 + \frac{\xi}{pr}\right)^{pr+\xi+\frac{1}{2}} e^{-(pr+\xi)}\sqrt{2\pi pr}$$

a podobně

$$(qr - \xi)! = (qr)^{qr-\xi}\left(1 - \frac{\xi}{qr}\right)^{qr-\xi+\frac{1}{2}} e^{-(qr-\xi)}\sqrt{2\pi rq},$$

takže po dosazení do (43) a jednoduché úpravě bude přibližně

$$y(\xi) = \frac{1}{\sqrt{2\pi rpq}}\left(1 + \frac{\xi}{pr}\right)^{-(pr+\xi+\frac{1}{2})}\left(1 - \frac{\xi}{qr}\right)^{-(qr-\xi+\frac{1}{2})}. \quad (44)$$

K osvětlení, jak se přibližuje tento výraz k (43), srovnáváme odchylky ξ od průměru se směrodatnou odchylkou $\sigma_x = \sqrt{rpq}$, která je řádu $\sqrt{r}$, není-li p ani q příliš malé. Musíme tedy předpokládati r tak velké, aby bylo možno zanedbati $\frac{\xi}{r}$, ale $\sqrt{\frac{\xi}{r}}$ musí míti takové konečné hodnoty, jaké se nám vyskytují, když posuzujeme odchylky ξ srovnáváním se směrodatnou odchylkou.

Můžeme tedy výraz (44), který napíšeme ve tvaru

$$y(\xi) = \frac{1}{\sqrt{2\pi rpq}} A \,.\, B,$$

zjednodušiti s uvedenou přibližností, neboť

$$\log A = -\left(rp + \xi + \frac{1}{2}\right)\left[\frac{\xi}{rp} - \frac{\xi^2}{2r^2p^2} + \frac{\xi^3}{r^3}\Phi_1(\xi)\right]$$

$$\log B = -\left(rq - \xi + \frac{1}{2}\right)\left[-\frac{\xi}{rq} - \frac{\xi^2}{2r^2q^2} - \frac{\xi^3}{r^3}\Phi_2(\xi)\right],$$

takže

$$\log y(\xi)\sqrt{2\pi rpq} = \frac{(p-q)\xi}{2rpq} - \frac{\xi^2}{2rpq} + \frac{\xi^2}{r^2}\Phi_3(\xi) =$$

$$= \frac{\xi^2}{2rpq} + \frac{\xi}{r}\Phi(\xi),$$

kde $\Phi_i(\xi)$ a $\Phi(\xi)$ jsou konečné, neboť představují součty konvergentních řad mocnin zlomku $\frac{\xi}{r}$, který je libovolně malý.

Je-li tudíž r tak velké, že $\frac{\xi}{r}\Phi(\xi)$ je malé, a tedy zanedbatelné, dostáváme

$$y(\xi) = \frac{1}{\sqrt{2\pi rpq}} e^{-\frac{\xi^2}{2\pi rpq}}.$$

Vzhledem k tomu, že $\sigma_x^2 = rpq$, můžeme také psáti

$$y(\xi) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma_x^2}}, \tag{45}$$

což je normální křivka rozdělení četností.

Ve směrodatné proměnné $\frac{x-\bar{x}}{\sigma_x} = \frac{\xi}{\sigma_x} = t$ má pak tvar

$$y(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}. \tag{46}$$

Křivka je symetrická podle průměru, do něhož jsme položili počátek souřadnic, který je tedy v bodě $t = 0$ a této

hodnotě odpovídá maximální pořadnice

$$y(0) = \frac{1}{\sqrt{2\pi}} = 0{,}39894.$$

Uvedeme si přehled několika pořadnic v intervalech $0{,}5\sigma_x$:

$\xi/\sigma_x =$	0,5	1,0	1,5	2,0	2,5	3,0
$y =$	0,35207	0,24117	0,12952	0,05399	0,01753	0,00443
$y/y(0) =$	0,88250	0,60653	0,32465	0,13534	0,04394	0,01111

Podrobnější tabulku poměru pořadnic $y : y(0)$ možno najíti na př. v [9]. Druhá derivace výrazu (45) je

$$\frac{\mathrm{d}^2 y}{\mathrm{d}\xi^2} = \frac{1}{\sqrt{2\pi}\sigma_x^3}\left(\frac{\xi^2}{\sigma_x^2} - 1\right) e^{-\frac{\xi^2}{2\sigma_x^2}},$$

z čehož plyne, že křivka má dva inflexní body pro $\xi = \pm \sigma_x$, neboť nejbližší derivace v těch bodech od nuly různá je třetí, tedy lichého stupně. Tečny v těchto bodech křivky protínají osu ξ v bodech $\xi = \pm 2\sigma_x$.

Momenty lichého řádu kolem průměru jsou pro symetrickou křivku Laplace-Gaussovu, jako pro každou symetrickou křivku, rovny nule, tedy $\mu_{x,1} = \mu_{x,3} = \mu_{x,5} = \ldots = 0$. Pro momenty sudého řádu lze odvoditi rekurentní vztah [6]

$$\mu_{x,2i} = (2i - 1)\,\sigma_x^2 \mu_{x,2i-2} \tag{47}$$

takže

$$\mu_{x,2} = \sigma_x^2, \qquad \mu_{x,4} = 3\sigma_x^4, \qquad \mu_{x,6} = 15\sigma_x^6, \ldots \tag{48}$$

Jako jsme viděli u histogramu, že celá jeho plocha představuje rozsah souboru, tak jej také zde znázorňuje plocha ohraničená křivkou a osou x. Tato plocha je dána při spojité proměnné integrálem

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}t^2}\,\mathrm{d}t = 1$$

vzhledem k tomu, že

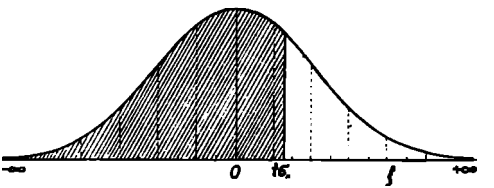
$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}t^2}\,\mathrm{d}t = \sqrt{2\pi}.$$

Máme-li soubor rozsahu r, pak je rovnice normální křivky

$$y(\xi) = \frac{r}{\sigma_x\sqrt{2\pi}}\, e^{-\frac{\xi^2}{2\sigma_x^2}} \tag{49}$$

a maximální pořadnice pro $\xi = 0$ je $y_r(0) = \dfrac{r}{\sigma_x\sqrt{2\pi}}$.

Vzhledem k souměrnosti křivky, je část ohraničená osou ξ



Obr. 14a. Úseky plochy normální křivky četností.

a křivkou v mezích od $-\infty$ do 0 rovna (viz obr. 14a) polovině celé plochy, tedy 0,5. Část plochy od $-\infty$ až do $\xi = t\sigma_x$, kde t je kladné číslo, bude

$$F(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{1}{2}t^2}\,\mathrm{d}t = 0{,}5 + \tfrac{1}{2}\alpha(t), \tag{50}$$

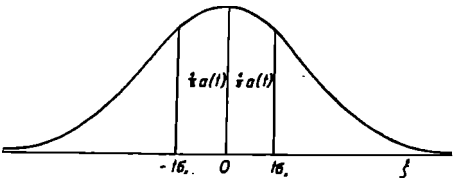
kde jsme zavedli

$$\tfrac{1}{2}\alpha(t) = \frac{1}{\sqrt{2\pi}}\int_{0}^{t} e^{-\frac{1}{2}t^2}\,\mathrm{d}t. \tag{51}$$

Bude tedy plocha pásu (obr. 14b) mezi $\xi = -t\sigma_x$ a $\xi = +t\sigma_x$

$$\alpha(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{1}{2}t^2} \, dt = \frac{2}{\sqrt{2\pi}} \int_{0}^{t} e^{-\frac{1}{2}t^2} \, dt. \tag{52}$$

Tyto hodnoty můžeme sestaviti do tabulky [6] pro různá t, t. j. pro různé hodnoty odchylky od průměru vyjádřené ve směrodatné odchylce jako jednotce. Tak je na př.



Obr. 14b.

$\frac{\xi}{\sigma_x} = t$	0,6745	1	$\sqrt{2}$	2	3
$\alpha(t)$	0,5	0,6827	0,8427	0,9545	0,9973
B.-Č.		0	0,5	0,750	0,889

Hodnota $\xi_p = 0{,}6745\sigma_x$ se nazývá také **pravděpodobná chyba**. Je patrno, že v mezích $\bar{x} \pm \xi_p$ je polovina celého souboru a tedy rovněž polovina vně těchto mezí. Je tudíž tato hodnota kvartilovou odchylkou. Hodnota $\xi_m = \sqrt{2}\sigma_x$ se nazývá **modul**. Užívá se jí také někdy za jednotku, v níž se vyjadřuje proměnná, takže $\frac{\xi}{\sigma_x\sqrt{2}} = \gamma$ a dostáváme pak funkci

$$\Phi(\gamma) = \frac{2}{\sqrt{\pi}} \int_{0}^{\gamma} e^{-\gamma^2} \, d\gamma, \tag{53}$$

která bývá tabelována. Přesvědčíme se ovšem snadno, že $\alpha(t) = \Phi(\gamma)$, provedeme-li substituci $t = \gamma\sqrt{2}$.

Ve třetím řádku byly pro srovnání uvedeny hodnoty vyplývající z teorému Bienaymé-Čebyševova.

Z uvedených čísel vidíme, jaké procento prvků souboru s normálním rozdělením četností je v určitých mezích odchylek od průměru.

Tak bude prvků:

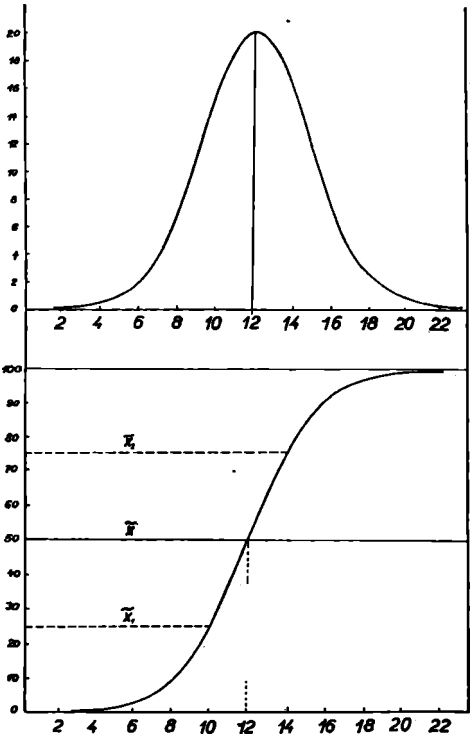
68,27%	s odchylkou	$\lvert\xi\rvert \leqq \sigma_x$,	ostatních je tedy	31,73%
95,45%	„	$\lvert\xi\rvert \leqq 2\sigma_x$,	„ „ „	4,55%
99,73%	„	$\lvert\xi\rvert \leqq 3\sigma_x$,	„ „ „	0,27%

V souboru s normálním rozdělením četností podle toho bude na př. 0,135% prvků s většími hodnotami než $\bar{x} + 3\sigma_x$ a rovněž tolik s menšími hodnotami než $\bar{x} - 3\sigma_x$, čili 0,27% prvků bude mimo interval $\pm 3\sigma_x$.

Vlastní praktický význam křivky Laplace-Gaussovy se jeví teprve při těchto daleko důležitějších otázkách, kde potřebujeme součet velkého počtu jednotlivých relativních četností, neboť méně nás zajímá otázka, jaká jest pravděpodobnost, že při 1200 vrzích kostkou padne právě $x = 180$krát šestka, jako spíše otázka, jaká je pravděpodobnost, že nebude odchylka od průměru $\bar{x} = 200$ větší než $200 - 180 = 20$. To vyžaduje zjistiti součet $\sum_{x=180}^{220} P_r(x)$ čili vypočítati podle Newtonovy formule (36) celkem 41 jednotlivých hodnot pravděpodobností $P_r(x)$ a sečísti. Integrací křivky Laplace-Gaussovy dosahuje se zde dalekosáhlého zjednodušení. Pro tento t. zv. Laplaceův integrál existují různé tabulky sestrojené pro různé argumenty; proto je třeba značné opatrnosti při jejich užívání a především řádného seznámení se s nimi.

Znázorníme si hodnoty funkce $F(t)$, probíhá-li proměnná t celý obor reálných čísel; dostáváme tak k normálnímu roz-

dělení četností součtovou křivku, která je znázorněna v obr. 15. Její pořadnice je v stupnici pětkrát zmenšené



Obr. 15. Součtová křivka k normálnímu rozdělení četností.

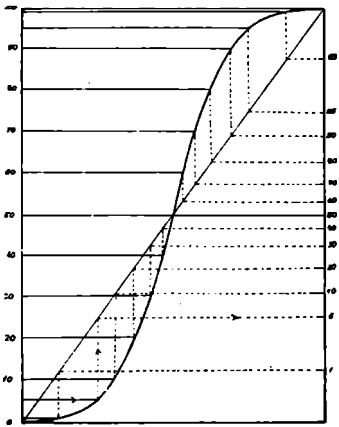
proti pořadnici příslušné normální křivky nahoře. Pomocí součtové křivky se snadno určuje, jak již víme, medián a kvartily.

(6,2) Normální rozdělení četností kvantitativního znaku. Odvodili jsme Laplace-Gaussovu křivku normálního rozdělení četností, pomocí náhodných výběrů, z nichž každý vykazuje určitou relativní četnost znaku $\frac{x}{r} = f$ a pochopili jsme tak vznik této křivky na základě binomického rozdělení četností. Normální rozdělení však vzniká také, provedeme-li mnohonásobné měření kvantitativního znaku na jednom předmětu (řada měření nějaké délky) nebo při měření určitého kvantitativního znaku na různých předmětech, jež jsou prvky jednoho statistického souboru (na př. délka listů určitého stromu). Pro výklad, jak vzniká normální rozdělení, v prvním případě si můžeme představit, že výsledek každého měření závisí na velikém počtu t. zv. elementárních příčin, z nichž každá je s to způsobiti nějakou elementární odchylku od skutečnosti. Tyto odchylky jsou v obou směrech stejně pravděpodobné a vzájemně nezávislé. Je to tedy analogie s náhodnými výběry koulí z osudí se stejnou pravděpodobností pro bílou i černou nebo analogie házení mincí. Takový výklad byl sestrojen původně pro teorii chyb při měření; přenášel se pak také na druhý případ, jímž se zabýváme ve statistice. Je však také jiný výklad, který snad lépe vystihuje skutečnost, takže se ho můžeme přidržeti. Vychází od hypotesy, že každá hodnota kvantitativního znaku je součtem množství neznámých a nezávislých sčítanců. Na př. délka nějakého předmětu (listu) se skládá z délky velkého množství nezávislých součástek (buněk). Také tudíž každá jednotlivá odchylka je součtem množství malých neznámých veličin, elementárních odchylek. Rozdělení těchto součtů je blízké normálnímu i když by rozdělení sčítanců nebylo normální. (Tak si můžeme vysvětlit, že se vyskytuje normální rozdělení četností také pro kvantitativní znak ve statistických souborech.)

(6,3) Pravděpodobnostní stupnice. Součtovou křivku patřící k normální křivce lze znázorniti přímkou, zvolíme-li

vhodnou stupnici pro pořadnici. Souvislost pravidelné stupnice relativních četností v procentech se stupnicí t. zv. pravděpodobnostní rovněž v procentech je vyznačena nomograficky v obr. 16, kde jsou patrny body přímky odpovídající bodům součtové křivky.

V této stupnici je znázorněna součtová křivka rozdělení četností na str. 29 (obr. 4b). Podle toho, jak se odchyluje od přímky, můžeme posouditi, že pozorované rozdělení četností se liší od normálního.



Obr. 16. Převod pravidelné stupnice na stupnici pravděpodobnostní.

Normální křivku lze rovněž převésti na přímku, zvolíme-li v pravoúhlé soustavě na ose úseček kvadratickou stupnici a na ose pořadnic logaritmickou stupnici [7, str. 19].

(6,4) Poissonovo rozdělení četností. (Exponenciela Poissonova.) Abychom z binomického rozdělení četností odvodili ještě jiné křivky rozdělení četností, budeme hledati pro funkci

$$y = \frac{r!}{x!\,(r-x)!}\,p^x q^{r-x}$$

vhodný výraz, který by ji vyjádřil přibližně v těch případech, kdy základní pravděpodobnost výskytu pozorovaného znaku p je malá, ale tak, že $rp = \lambda$ je číslo konečné pro libovolně veliké r.

Především je

$$\frac{r!}{(r-x)!} = r\,(r-1)\,(r-2)\ldots(r-x+1).$$

Dále pišme

$$p = \frac{\lambda}{r} \text{ a tedy } q = 1 - \frac{\lambda}{r},$$

takže bude tedy

$$y = \left(1 - \frac{1}{r}\right)\left(1 - \frac{2}{r}\right) \ldots \left(1 - \frac{x-1}{r}\right) \cdot \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{r}\right)^r q^{-x}.$$

Přibližný výraz dostaneme pro velká r, zanedbáme-li veličiny řádu $\frac{1}{r}$, takže především součin prvních $x - 1$ činitelů v závorkách, který je mezi 1 a $1 - \frac{x(x-1)}{2r}$, položíme roven přibližně 1.

Dále můžeme místo $\left(1 - \frac{\lambda}{r}\right)^r$ klásti přibližně $e^{-\lambda}$, což je limita, k níž výraz spěje pro $r \to \infty$. Konečně q^{-x} spěje pro velká r k 1, neboť $q^{-x} = \left[\left(1 - \frac{\lambda}{r}\right)^{-r}\right]^{\frac{x}{r}}$, což spěje k $(e^{\lambda})^{\frac{x}{r}}$, a tedy pro velká r se $\frac{x}{r}$ blíží k nule. Z toho všeho tudíž vyplývá, že můžeme přibližně klásti

$$y = \frac{e^{-\lambda}\lambda^x}{x!}; \tag{54}$$

tento výraz se obyčejně označuje symbolem $\psi(x)$ a nazývá se exponenciela Poissonova, udávající pravděpodobnost, že se vyskytuje x-krát pozorovaný znak, který patří mezi tak zv. řídké jevy, jejichž pravděpodobnost p je malá. Bortkiewicz jej nazval zákonem malých čísel.

Pravděpodobnosti, že se objeví pozorovaný znak právě 0, 1, 2, ... krát, jsou dány jednotlivými členy řady

$$e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right).$$

Ačkoliv jsme předpokládali při odvozování Poissonovy exponenciely, že x je malé vzhledem k r, dostáváme k rozdělení četností, vyjádřenému touto exponencielou, klademe-li za x všechna celá čísla od $x = 0$ do $x = r$, jednoduché a důležité výsledky pro průměr a směrodatnou odchylku. Pro velká r platí přibližně

$$e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^r}{r!}\right) = 1, \tag{55}$$

neboť součet v závorce je přibližně roven e^{λ}. Jednotlivé členy pravé strany jsou tedy relativní četnosti. Vynásobíme-li každou z nich příslušnou hodnotou znaku 0, 1, 2, ..., r, dostaneme průměr

$$\begin{aligned}\bar{x} &= e^{-\lambda}\left(0 + \lambda + \lambda^2 + \frac{\lambda^3}{2!} + \ldots + \frac{\lambda^r}{(r-1)!}\right) = \\ &= \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^{r-1}}{(r-1)!}\right) = \lambda,\end{aligned} \tag{56}$$

neboť součet v závorce je pro velká r přibližně týž jako v rovnici (55).

Podobně dostaneme pro druhý moment obecný

$$\mu'_{x,2} = e^{-\lambda}\left(0 + \lambda + 2\lambda^2 + \frac{3\lambda^3}{2!} + \ldots + \frac{r\lambda^r}{(r-1)!}\right),$$

takže rozptyl

$$\mu_{x,2} = \mu'_{x,2} - \bar{x}^2$$

bude

$$\mu_{x,2} = \lambda e^{-\lambda}\left(1 + 2\lambda + \frac{3\lambda^2}{2!} + \ldots + \frac{r\lambda^{r-1}}{(r-1)!}\right) - \lambda^2,$$

což lze také psáti

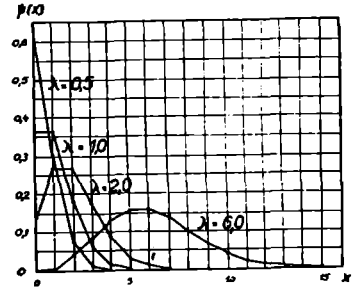
$$\begin{aligned}\mu_{x,2} = \lambda e^{-\lambda}&\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^{r-1}}{(r-1)!}\right) + \\ &+ \lambda^2 e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^{r-2}}{(r-2)!}\right) - \lambda^2;\end{aligned}$$

vidíme tedy vzhledem k (55)

$$\sigma_x^2 = \lambda + \lambda^2 - \lambda^2 = \lambda. \qquad (57)$$

Je tedy rozptyl roven průměru. Vzhledem k tomu, že $\lambda = rp$, je to tedy hodnota blízká rpq, kterou jsme našli pro normální rozdělení četností, neboť q se liší velmi málo od 1.

Hodnoty Poissonovy exponenciální limity (54) byly tabelovány pro různá λ a x; lze je najíti na př. v tabulkách [8]. Průběh jejich je znázorněn na obr. 17 pro $\lambda = 0{,}5$, 1, 2, 6. Je jasně viděti, že od úplné nesouměrnosti přecházejí křivky pro rostoucí λ k tvaru stále souměrnějšímu.



Obr. 17. Exponenciela Poissonova.

(6,5) Pearsonův systém křivek četnosti. Viděli jsme, že lze odvoditi z binomického rozdělení četností čili z formule Newtonovy (36) celý systém křivek rozdělení četnosti. Mohou však býti odvozeny ještě obecnější systémy. Představme si, že základní soubor konečného rozsahu N obsahuje k prvků, majících pozorovaný alternativní znak a $N - k$ prvků, které jej nemají. Vyjmeme-li z tohoto základního souboru částečné soubory o rozsahu r prvků, můžeme tak učiniti celkem $\binom{N}{r}$ různými způsoby, čili můžeme dostati tolik různých výběrů. Každý z těchto výběrů má určitý počet x prvků s uvažovaným znakem. Kladné celé číslo x je v intervalu od 0 do r, když předpokládáme $k \geqq r$. Abychom stanovili, kolik může býti různých výběrů, jež mají určitý počet x prvků s pozorovaným znakem, uvědomíme si, že je celkem $\binom{k}{x}$ skupin, jež obsahují x různých

prvků z daných k prvků s uvažovaným znakem v základním souboru, a ke každé z těchto skupin lze přiřaditi $\binom{N-k}{r-x}$ různých skupin tvořených ze zbývajících $r-x$ prvků, které nemají uvažovaný znak a doplňují skupinu na celkový rozsah výběru r. Vidíme, že tedy bude hledaný počet různých výběrů čili absolutní četnost $\binom{k}{x}\binom{N-k}{r-x}$. Relativní četnost jejich dostaneme, dělíme-li poslední výraz celkovým počtem různých možných výběrů rozsahu r, takže bude vyjádřena funkcí

$$f(x) = \frac{1}{\binom{N}{r}}\binom{k}{x}\binom{N-k}{r-x} \tag{58}$$

pro hodnoty $x = 0, 1, 2, \ldots, r$; jsou to postupně za sebou jdoucí členy konečné řady hypergeometrické. Z této funkce vyšel K. Pearson, aby odvodil t. zv. Pearsonův systém křivek rozdělení četností, v němž jsou křivky (45) a (54) zahrnuty jako zvláštní případy [6], neboť také binomická funkce (36) je zvláštním případem hypergeometrické, která v ni přechází pro nekonečný rozsah základního souboru, takže $N = \infty$, $k = \infty$, ale jejich poměr $\frac{k}{N} = p$ je konstantní a konečný.

Vhodnou volbou typu křivky je pak možno s postačujícím přiblížením vyjádřiti statisticky pozorovaná rozdělení četností. Tato volba je tu usnadněna tím, že bylo odvozeno — pomocí momentů — kriterium, které umožňuje rozhodnouti se mezi možnými typy pro vhodnější.

(6,6) Pólyovo výběrové schema pro jevy vázané. K určité funkci hypergeometrické jsme vedeni, provádíme-li ze základního souboru o N prvcích, z nichž má k prvků pozorovaný znak, výběr rozsahu r tak, že vyjmeme prvek a zjistíme, má-li pozorovaný znak. V kladném případě se

počet prvků s pozorovaným znakem v základním souboru zvětší o $1 + \Delta$; neměl-li prvek pozorovaný znak, zvětší se o $1 + \Delta$ počet těchto druhých prvků v základním souboru. V okamžiku, když jsme vyňali r prvků, bude míti základní soubor celkem $N + r\Delta$ prvků.

Bylo-li mezi nimi x prvků s pozorovaným znakem a tudíž $r - x$ ostatních, je v základním souboru $k + x\Delta$ prvků s pozorovaným znakem a $N - k + (r - x)\Delta$ ostatních. Pravděpodobnost výskytu pozorovaného znaku je na začátku v základním souboru $\frac{k}{N} = p$ a opačná $\frac{N-k}{N} = q$; mění se po vynětí každého prvku do výběru, takže po vynětí r-tého je $\frac{k + x\Delta}{N + r\Delta}$ resp. $\frac{N - k + (r - x)\Delta}{N + r\Delta}$.

Pravděpodobnost, že prvních x prvků bude míti pozorovaný znak ve výběru rozsahu r, bude jako složená pravděpodobnost dána součinem

$$\frac{k}{N} \cdot \frac{k+\Delta}{N+\Delta} \cdot \ldots \cdot \frac{k + (x-1)\Delta}{N + (x-1)\Delta} \cdot \frac{N-k}{N+x\Delta} \cdot \frac{N-k+\Delta}{N+(x+1)\Delta} \cdot \ldots \cdot \frac{N-k+(r-x-1)\Delta}{N+(r-1)\Delta}.$$

Zavedeme-li označení $\frac{\Delta}{N} = \delta$, přechází poslední výraz na tvar

$$\frac{p}{1} \cdot \frac{p+\delta}{1+\delta} \cdot \ldots \cdot \frac{p + (x-1)\delta}{1 + (x-1)\delta} \cdot \frac{q}{1+x\delta} \cdot \frac{q+\delta}{1+(x+1)\delta} \cdot \ldots \cdot \frac{q + (r-x-1)\delta}{1 + (r-1)\delta}.$$

Pravděpodobnost, že bude ve výběru téhož rozsahu r jiných x prvků se znakem pozorovaným, bude dána týmž výrazem, jen pořadí jednotlivých faktorů bude jiné.

Kombinací, v nichž se může vyskytnouti mezi r prvky x s pozorovaným znakem je $\binom{r}{x}$, takže celkem pravděpodob-

nost, že mezi r prvky výběru provedeného ze základního souboru, který se uvedeným způsobem mění, bude x prvků s pozorovaným znakem, je

$$f(x,r) = \binom{r}{x} \cdot \frac{p\,(p+\delta)\ldots[p+(x-1)\,\delta]\,q\,[q+\delta]\ldots[q+(r-x-1)\,\delta]}{[1+\delta]\,[1+2\delta]\ldots[1+(r-1)\,\delta]}.$$

Je-li pravděpodobnost p malá, ale pro velká r je $rp = \lambda$ konečné číslo, a při kladném δ označíme $r\delta = d > 0$, platí přibližně

$$f(x,r) = \frac{1}{x!}\,\lambda\,(\lambda+d)\,(\lambda+2d)\ldots(\lambda+\overline{x-1}\,d)\,(1+d)^{-\frac{\lambda}{d}-x} \tag{59}$$

což se nazývá zákonem Pólyovým a je zobecněním exponenciely Poissonovy (54), která z něho vyplývá jako limita pro $d = 0$.

Pro uvedené schema výběrové to znamená, že $\Delta = 0$ čili redukuje se na případ schematu Bernoulliho o konstantní pravděpodobnosti p.

Jiný případ dostaneme pro $\Delta = -1$, který znamená, že prvek se vyjme do výběru a rozsah základního souboru se tím vždy o jeden prvek zmenšuje, což je případ Pearsonův, odpovídající konečnému souboru základnímu, do něhož se vyňatý prvek nevrací zpět.

Zákon Pólyův uvádíme vzhledem k jeho obecnosti a také proto, že se osvědčil k vystižení případů, kde se nejedná o jevy nezávislé, nýbrž nějakým způsobem vázané, jako je případ úmrtnosti vlivem nakažlivých nemocí, nebo smrti cestujících následkem neštěstí na drahách a pod.

(6,7) Rozvoje v řady. (Řada Brunsova.) Praktický problém vyjádření pozorovaného rozdělení četností analyticky je také velmi obecně řešen pomocí řady, jejímž prvním členem je funkce Laplace-Gaussova jako vytvořující a dalšími členy její derivace.

Omezíme-li se jen na první dva členy, od nuly různé, dostáváme vyjádření

$$f(\xi) = \varphi(\xi) - \varphi(\xi)\,\frac{\mu_{x,3}}{3!\,\sigma_x^3}\left(\frac{3\xi}{\sigma_x} - \frac{\xi^3}{\sigma_x^3}\right), \tag{60}$$

kde

$$\varphi(\xi) = \frac{1}{\sigma_x\sqrt{2\pi}}\,e^{-\frac{\xi^2}{2\sigma_x^2}}, \quad \varphi(\xi)\left(\frac{3\xi}{\sigma_x} - \frac{\xi^3}{\sigma_x^3}\right) = \varphi'''(\xi),$$

takže z pozorovaného rozdělení četností musíme stanoviti první tři momenty, abychom určili potřebné tři konstanty, jež se ve výrazu vyskytují $\bar{x}$, σ_x, $\mu_{x,3}$.

(Řada Poisson-Charlierova.) Také jiné funkce mohou sloužiti k podobným rozvojům v řadu. Zvláště jednoduchý a pro vystižení nesymetrických rozdělení četností vhodný je rozvoj pomocí exponenciely Poissonovy, který uvedeme také bez odvozování

$$f(x) = \psi(x) + \tfrac{1}{2}(\mu_{x,2} - \lambda)\,\Delta^2\,\psi(x). \tag{61}$$

$\psi(x)$ značí exponencielu Poissonovu (54) a druhá diference je

$$\Delta^2\,\psi(x) = \psi(x) - 2\psi\,(x-1) + \psi\,(x-2).$$

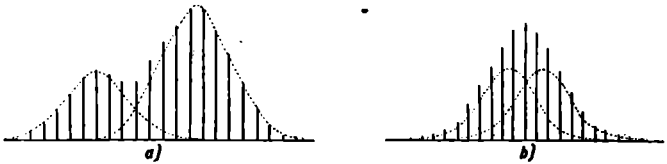
Na rozdíl od Pearsonova systému křivek nemáme zde kriteria, které by nám pomohlo rozhodnouti, zda máme použíti řady Brunsovy nebo Poisson-Charlierovy pro rozpojitou proměnnou x, takže musí rozhodnouti statistik sám podle vhodnosti a účelnosti.

Praktický význam rozvojů vytvořených pomocí jiných funkcí je omezen požadavkem rychlé konvergence řady, aby bylo možno se omeziti jen na několik málo členů.

(6,8) Vícevrcholová rozdělení četností. Někdy se vyskytují soubory, jejichž rozdělení četností má dva vrcholy (čili dvě maxima) jako v obr. 18a nebo více vrcholů. Vznik takového rozdělení četností se vysvětluje tím, že soubor zahrnuje prvky nestejnorodé podle některého znaku, takže bychom dostali dvě různá rozdělení četností, kdybychom

podle něho soubor roztřídili. Představujeme si tedy, že výsledná křivka rozdělení četností vznikla superposicí dvou jednoduchých křivek; při tom by ovšem mohla vzniknouti také křivka jednovrcholová (obr. 18b).

Za účelem oddělení obou jednoduchých křivek je možno použíti pro některé tvary dvojvrcholových rozdělení čet-



Obr. 18a, b. Dvojvrcholové rozdělení četností.

ností křivek normálních, takže pak dané rozdělení je vyjádřeno rovnicí

$$r\,f(x) = \frac{r_1}{{}_1\sigma_x\sqrt{2\pi}}\,e^{-\frac{(x-\bar{x}_1)^2}{2\,{}_1\sigma_x^2}} + \frac{r_2}{{}_2\sigma_x\sqrt{2\pi}}\,e^{-\frac{(x-\bar{x}_2)^2}{2\,{}_2\sigma_x^2}},$$

kde čísla r_1 a r_2 udávají, v jakém poměru se vyskytují v celkovém rozsahu r prvky prvního a druhého souboru složkového. Konstanty vypočítáme pomocí momentů celkového rozdělení četností.

Úloha: Vypočítejte konstanty pro jednoduchý případ, kdy oba vrcholy spadají do téhož místa, takže $\bar{x}_1 = \bar{x}_2 = \mu'_{x,1}$.

V tomto případě je rozdělení symetrické, takže momenty lichého stupně kolem průměru se rovnají nule, tedy $\mu_{x,1} = \mu_{x,3} = \mu_{x,5} = 0$ a ostatní jsou vzhledem k rovnicím (48) postupně

$$\begin{aligned} r &= r_1 + r_2 \\ r\mu_{x,2} &= r_1\,{}_1\sigma_x^2 + r_2\,{}_2\sigma_x^2 \\ r\mu_{x,4} &= 3(r_1\,{}_1\sigma_x^4 + r_2\,{}_2\sigma_x^4) \\ r\mu_{x,6} &= 15(r_1\,{}_1\sigma_x^6 + r_2\,{}_2\sigma_x^6). \end{aligned}$$

Řešením těchto čtyř rovnic je možno určiti r_1, r_2, ${}_1\sigma_x$, ${}_2\sigma_x$, neboť ${}_1\sigma_x{}^2$ a ${}_2\sigma_x{}^2$ dostaneme jako dva kořeny jedné rovnice druhého stupně.

(6,9) Příklady.

1. (Normální rozdělení četností.) Vyjádříme skupinové rozdělení četností dané v sloupci (1) a (2) pomocí křivky Laplace-Gaussovy. Použijeme k tomu částí její plochy, daných výrazem $F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}t^2}\, dt$, kde $t = \frac{x - \bar{x}}{\sigma_x} = \frac{u - \bar{u}}{\sigma_u}$ vzhledem k (15) a (17).

x_i	n_i	u_i	$u_i n_i$	$u_i{}^2 n_i$	$u_i - \bar{u}$
(1)	(2)	(3)	(4)	(5)	(6)
17	14	—3	— 42	126	—2,095
22	121	—2	—242	484	—1,095
27	335	—1	—335	335	—0,095
32	349	0	0	0	0,905
37	150	1	150	150	1,905
42	29	2	58	116	2,905
47	2	3	6	18	∞
Σ	1000		—405	1229	

$t = \frac{u_i - \bar{u}}{{}_0\sigma_u}$	$F(t)$	$\Delta F(t)$	$r \,.\, \Delta F(t)$
(7)	(8)	(9)	(10)
—2,1140	1—0,9827	0,0173	17,3
—1,1049	1—0,8654	0,1173	117,3
—0,0959	1—0,5382	0,3272	327,2
0,9132	0,8195	0,3577	357,7
1,9223	0,9727	0,1532	153,2
2,9314	0,9983	0,0256	25,6
∞	1,0000	0,0017	1,7
		1,0000	1000,0

$\bar{u} = -0,405$ $\qquad \bar{x} = 29,975$

$\mu'_{u,2} = 1,229$ $\qquad {}_0\mu_{u,2} = 0,982$

$\mu_{u,2} = 1,065$ $\qquad {}_0\sigma_u = 0,991$

Výsledek v sl. 10. dává t. zv. teoretické rozdělení četností.

2. (Poissonova exponenciela.) Počet úmrtí žen starších 85 let, pozorovaný denně v období tří let je uveden v (1) a (2) sloupci tabulky rozdělení četností. Vzhledem k jeho nesymetrickému tvaru se pokusíme o vyjádření exponencielou Poissonovou. Potřebujeme k tomu cíli zjistit průměr rozdělení λ. Celkový rozsah je $r = 1086$.

Počet úmrtí denně	Počet dní				
x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$\psi(x)$	$r\,\psi(x)$
(1)	(2)	(3)	(4)	(5)	(6)
0	364	0	0	0,30360	329,7
1	376	376	376	0,36179	392,9
2	218	436	872	0,21568	234,2
3	89	267	801	0,08576	93,1
4	33	132	528	0,02559	27,8
5	13	65	325	0,00611	6,7
6	2	12	72	0,00122	1,3
7	1	7	49	0,00025	0,3
Σ	1086	1295	3023	1,00000	1086,0

$$\lambda = \bar{x} = 1{,}1924$$
$$\mu'_{x,2} = 2{,}7836$$
$$\mu_{x,2} = 1{,}3618$$

3. (Řada Poisson-Charlierova.) Pryskyřník (Ranunculus) je rostlina s korunou zpravidla pětičetnou, vyskytují se však i případy s korunou vícečetnou. Pozorováním 222 květů bylo zjištěno dole uvedené rozdělení četností. Pro jeho analytické vyjádření použijeme řady Poisson-Charlierovy (61). Potřebujeme vyčíslit koeficient druhého členu a další postup je patrný z tabulky výpočtů.

$$\lambda = \bar{u} = 0{,}631$$
$$\mu'_{u,2} = 1{,}315$$
$$\mu_{u,2} = 0{,}917$$
$$\tfrac{1}{2}(\mu_{x,2} - \lambda) = c = 0{,}143$$

Počet lístků v koruně	Četnost				
x_i	n_i	u_i	$u_i n_i$	$u_i^2 n_i$	$\psi(u)$
(1)	(2)	(3)	(4)	(5)	(6)
5	133	0	0	0	0,53262
6	55	1	55	55	0,33497
7	23	2	46	92	0,10588
8	7	3	21	63	0,02243
9	2	4	8	32	0,00359
10	2	5	10	50	0,00051
Σ	222		140	292	1,00000

$\Delta\,\psi(u)$	$\Delta^2\,\psi(u)$	$c\Delta^2\,\psi(u)$	y	$r\,.\,y$
(7)	(8)	(9)	(10)	(11)
0,53262	0,53262	0,07616	0,6088	135,2
—0,19765	—0,73027	—0,10443	0,2305	51,2
—0,22909	—0,03144	—0,00450	0,1014	22,5
—0,08345	0,14564	0,02083	0,0433	9,6
—0,01884	0,06461	0,00924	0,0128	2,8
—0,00308	0,01576	0,00225	0,0028	0,7
				222,0

Ve sloupci (6) určena hodnota $\psi(u)$ pro znak $u \geqq 5$, aby se docílilo součtu relativních četností 1; diference pak byly počítány tak, jakoby to byla hodnota ψ (5), neboť vliv této úpravy je zanedbatelný.

Úloha: Vyjádřete pomocí řady Poisson-Charlierovy rozdělení četností použité v předchozím příkladu.

(7,1) Aplikace a zobecnění Bernoulliova teorému. (Od Bernoulliova teorému k závěrům o skutečném průběhu jevů.) Laplaceovým integrálem jsme získali důležitý prostředek k řešení některých úloh praktické statistiky, jež se často opakují. Proto umožňuje statistikovi ohromnou úsporu práce a času. Nesmíme však nikdy zapomínati, že pro konečný rozsah souboru r znamená jen přibližnou formuli, jejíž meze chyb nelze obyčejně ani dosti přesně odhadnout. Především můžeme vhodně použít Laplaceova integrálu k takové formulaci Bernoulliova teorému, jež by usnadnila jeho praktickou aplikaci. Základní problém

Bernoulliův jest v určení pravděpodobnosti $P_r(x)$, že v náhodném výběru rozsahu r bude právě x prvků s pozorovaným znakem alternativním je-li p jeho relativní četnost v základním souboru. Tato pravděpodobnost je určena Newtonovou formulí (36). Můžeme pak snadno určit pravděpodobnost, že četnost x znaku ve výběru ze základního souboru o konstantní relativní četnosti p se odchýlí od průměru rp nejvýše o $\pm \xi_0$. Pro dosti velká r je hledaná pravděpodobnost

$$P_r(\bar{x} - \xi_0, \bar{x} + \xi_0)$$

dána Laplaceovým integrálem

$$\Phi(\gamma_0) = \frac{2}{\sqrt{\pi}} \int_0^{\gamma_0} e^{-\gamma^2} d\gamma \quad \text{pro} \quad \gamma_0 = \frac{\xi_0}{\sqrt{2rpq}}.$$

Bernoulliův teorém dostaneme z toho malou změnou proměnné. Uvažujeme místo četnosti x znaku jeho relativní četnost $f = \frac{x}{r}$ a ptáme se, jaká je pravděpodobnost, že relativní četnost znaku ve výběru má určitou hodnotu f. Označíme-li tuto pravděpodobnost $P_r(f)$, bude zřejmě $P_r(f) = P_r(x)$. Dále je $\frac{\bar{x} - \xi_0}{r} = \frac{rp - \xi_0}{r} = p - z_0$ píšeme-li $z_0 = \frac{\xi_0}{r}$. Můžeme tedy udati pravděpodobnost, že relativní četnost znaku ve výběru se bude odchylovati od relativní četnosti v základním souboru nejvýše o z_0, neboť je opět dána Laplaceovým integrálem

$$P_r\left(\frac{\bar{x} - \xi_0}{r}, \frac{x + \xi_0}{r}\right) = P_r(p - z_0, p + z_0) = \Phi\left(\frac{rz_0}{\sqrt{2rpq}}\right)$$

čili

$$P_r(p - z_0, p + z_0) = \Phi\left(z_0 \sqrt{\frac{r}{2pq}}\right). \tag{62}$$

Na pravé straně této rovnice je funkce z_0, která s rostoucím r spěje k 1 při každé hodnotě z_0, neboť $\Phi(\infty) = 1$. Platí tedy pro pevné z_0

$$\lim_{r\to\infty} P_r\,(p - z_0,\, p + z_0) = 1, \tag{63}$$

což je výrazem Bernoulliova teorému, který znovu vyslovíme: Je-li rozsah r náhodného výběru dosti velký, je pravděpodobnost, že relativní četnost alternativního znaku v něm se odchýlí od své relativní četnosti v základním souboru o méně než z_0 libovolně blízka 1, ať je z_0 jakkoliv malé.

Pravděpodobnosti velmi blízké 1 se také říká méně přesně „skoro-jistota". Potom může předcházející věta znít: Je skoro jisto, že relativní četnost bude libovolně blízko statistické pravděpodobnosti, je-li jen r dosti velké.

Je důležito si uvědomiti, že vývody až potud byly provedeny jen matematickými úvahami z oboru t. zv. kombinatoriky. Proto nemůžeme za tohoto stavu nic říci o tom, jaká četnost znaku c by se ve skutečnosti objevila, kdybychom vzali r prvků ze základního souboru rozsahu N. Mohli bychom dostati relativní četnost $\frac{r}{r} = 1$, kdybychom vzali prvky výběru z jedné části základního souboru, která má jen prvky se znakem c. Kdyby však byly všechny prvky se znakem c v jiné části základního souboru, dostali bychom při téže relativní četnosti p v základním souboru výsledek $\frac{0}{r} = 0$ při platnosti všech formulí, jež jsme si odvodili. Abychom mohli se svými vývody pokročiti k nějakým závěrům o vztahu mezi $\frac{x}{r}$ a p museli jsme udělati dodatečný předpoklad, že základní soubor rozsahu N je dobře promíchán čili prvky se znakem c jsou v něm více méně stejnoměrně rozděleny. Tomu promíchání jsme rozuměli technicky, tedy asi tak jako vznikne beton pečlivým promícháním cementu, štěrku, písku a vody. Podati pro pojem dobrého

promíchání ryze matematickou definici je ovšem úkol zcela jiný. Byly takové definice sestrojeny a založen na nich celý počet pravděpodobnosti. Tak na př. Misesova definice vychází od základního souboru nekonečného rozsahu zvaného kolektiv, který má tři vlastnosti: 1. Prvky tvoří posloupnost nepravidelnou. 2. Relativní četnost f_i znaku c spěje při neomezeném počtu prvků k pevné mezní hodnotě p; předpokládá se tedy, že existuje limita $\lim_{i=\infty} f_i = p$, zvaná pravděpodobnost. Říkáme, že relativní četnost f_i spěje ku p stochasticky a tato konvergence ve smyslu teorie pravděpodobnosti čili stochastická je charakterisována větou Bienaymé-Čebyševovou a Bernoulliovou. 3. V každé posloupnosti libovolně odvozené ze základní, je táž mezní hodnota relativní četnosti.

Tím, že v definici pravděpodobnosti předpokládáme nějakou limitu relativní četnosti, idealisujeme pozorovanou skutečnost k účelu definice. V některých směrech je to analogická idealisace jako přímka nebo koule v geometrii či hmota a síla ve fysice.

Přesvědčili jsme se, že lze jen matematickou cestou dospěti k závěru, že relativní četnost výběrů čili jednotlivých kombinací r-té třídy bude tím menší, čím je počet x prvků se znakem c vzdálenější od průměru $\bar{x} = rp$. Dokonce můžeme snadno pomocí integrálu Laplaceova uvésti čísla udávající, že celková relativní četnost výběrů, u nichž je odchylka $\xi = x - rp$ v mezích $\pm 2\sigma_x$ je rovna 0,9545 a tedy celková relativní četnost těch případů, v nichž ξ je menší než $-2\sigma_x$ a v nichž je větší než $+2\sigma_x$ je rovna $1 - 0{,}9545 = 0{,}0455$. Relativní četnost případů, v nichž je

$|\xi| > 3\sigma_x$ je 0,0027, pro $|\xi| > 4\sigma_x$ je 0,000063 atd.

Od těchto ryze matematických výsledků nás převádí ke skutečnostem světa nás obklopujícího věta, která je výsledkem našich zkušeností a možno říci skoro axiomem denního života, t. zv. věta Cournotova. Tato věta konsta-

tuje, že zřídka se stane, abychom vyňali náhodně (t. j. bez zvláštního vybírání a hledání) z promíchaného základního souboru prvek se znakem, jehož relativní četnost v něm, čili pravděpodobnost, je velmi malá. Kdyby v množství 10 000 zrnek hrachu bylo jedno černé, pak zřídka vyjmeme náhodně z promíchaného množství právě to černé zrnko.

Odvodili jsme pak, že větší odchylky četnosti x $\left(\text{resp. relativní četnosti } \frac{x}{r}\right)$ v náhodných výběrech od její hodnoty rp (resp. p) v základním souboru mají při dostatečně velkém r velmi malou relativní četnost v souboru rozsahu $\binom{N}{r}$ všech možných výběrů rozsahu r ze základního souboru rozsahu N. Soudíme tudíž, že ve statistických souborech dostatečně velkého rozsahu se vyskytují také ve skutečnosti jen velmi zřídka větší odchylky relativních četností od odpovídající jim statistické pravděpodobnosti. Tato věta bývá nazývána zákonem velkých čísel a je jednou z těch, které vyjadřují princip velkých čísel.

Naše vědění spočívající na principu velkých čísel není prosto jisté subjektivní libovůle. Vidíme to, chceme-li říci, do které hodnoty máme považovati relativní četnost za „velmi malou", takže pozorovaný znak se ve skutečnosti objeví jen „velmi zřídka" nebo jakou hodnotu relativní četnosti máme nejvýše připustiti, abychom mohli očekávati, že se pozorovaný znak „prakticky neobjeví". Rozhodnutí nezávisí jen na absolutní velikosti pravděpodobnosti a názoru badatele, nýbrž také na stupni důležitosti, jakou by pro něho subjektivně mohl mít nepravděpodobný jev, kdyby přece nastal.

V praksi se ustálila zvyklost, že za mez pravděpodobností, na které se ještě béře zřetel, se volí celková pravděpodobnost odchylek, které přesahují $\pm 3\sigma_x$. Našli jsme, že pravděpodobnost odchylek větších než trojnásobek směrodatné

odchylky je dána číslem $0{,}0027 = \frac{1}{370}$. Toto „pravidlo tří sigma" je nyní velmi populární. Kdyby ovšem závisel náš vlastní život na vyskytnutí se jevu, který má tuto pravděpodobnost 0,3%, nezdála by se nám jistě úplně zanedbatelnou. Zavádí se také v poslední době jistý decimální systém mezních pravděpodobností a sice 5% pro optimisty, 2%, a 1% pro pesimisty.

(7,2) Poissonovo zobecnění teorému Bernoulliova. Uvažujme výběr prvků se znakem alternativním, které mají různé základní pravděpodobnosti. Máme tedy r základních souborů, které mají relativní četnosti pozorovaného znaku c a doplňky na jednotku postupně $p_1, q_1;\ p_2, q_2;\ \ldots;\ p_r, q_r$. Zobecnění vztahu (62), které podal Poisson (1837), spočívá v tom, že se vezme z každého z těchto základních souborů jeden prvek a určí se pravděpodobnost $P_r(x)$, že výběr bude mít x prvků se znakem c. Klademe-li zase

$$x = rf, \text{ je } P_r(x) = P_r(f),$$

kde $P_r(f)$ značí pravděpodobnost, že dostaneme výběr rozsahu r s relativní četností f. Poisson dokázal, že také pro toto rozdělení $P_r(f)$ platí vztah (63) o mezní hodnotě, v němž p musí býti průměr čísel $p_1, p_2, \ldots, p_r$, takže v Laplaceově integrálu bude hranice

$$\gamma_0 = \frac{z_0 r}{\sqrt{2\sum_{k=1}^{r} p_k (1 - p_k)}}.$$

(7,3) Průměr a rozptyl rozdělení četností vzniklého tvořením součtů z několika náhodných proměnných. — (Bernoulliův problém jako zvláštní případ.) V jednom základním souboru jsou hodnoty, jichž nabývá kvantitativní znak, označený čísly $x_1, x_2, \ldots, x_l$, jimž odpovídá rozdělení relativních četností $p_1(x_1), p_1(x_2), \ldots, p_1(x_l)$, takže $\sum_{i=1}^{l} p_1(x_i) = 1$. Tak se stává znak náhodnou proměn-

nou. Průměr je podle definice

$$x_1 \cdot p_1(x_1) + x_2(p_1(x_2) + \ldots + x_l p_1(x_2) = \sum_{i=1}^{l} x_i p_1(x_i). \quad (64)$$

Tato hodnota, jakožto parametr základního souboru se označuje často zvláštním symbolem $\mathfrak{E}(x)$, jehož jsme již užili; je obdobným na př. znaménku integračnímu a odlišuje se tím jasně od průměru jako charakteristiky výběrové. Budeme ji nazývati očekávaná hodnota; vyskytují se v počtu pravděpodobnosti také názvy střední hodnota nebo matematická naděje. V druhém základním souboru buďtež hodnoty kvantitativního znaku $y_1, y_2, \ldots, y_m$ s rozdělením četností $p_2(y_1), p_2(y_2), \ldots, p_2(y_m)$. Očekávaná hodnota této náhodné proměnné je

$$\mathfrak{E}(y) = \sum_{i=1}^{m} y_i p_2(y_i). \quad (65)$$

Odvodíme nyní očekávanou hodnotu součtu dvou náhodných proměnných. Vezmeme náhodně z prvního základního souboru jeden prvek. Pravděpodobnost, že hodnota znaku bude x_i je $p_1(x_i)$. Obdobně bude $p_2(y_k)$ pravděpodobnost, že z druhého základního souboru vezmeme náhodně y_k. Pravděpodobnost, že se současně vyskytne znak x_i a y_k, je podle pravidla o složené pravděpodobnosti dána součinem $p_1(x_i)\, p_2(y_k) = p_{ik}$, a to je tedy také pravděpodobnost, že dostaneme určitý součet $x_i + y_k$.

Pro součet obou náhodných proměnných chceme najíti očekávanou hodnotu jako průměr. Sestavíme si tedy hodnoty pravděpodobností čili relativních četností v nově vzniklém základním souboru, pro jednotlivé možné páry hodnot x_i a y_k do této tabulky

	x_1	x_2	$\ldots$	x_l
y_1	p_{11}	p_{21}	$\cdots$	p_{l1}
y_2	p_{12}	p_{22}	$\cdots$	p_{l2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_m	p_{1m}	p_{2m}	$\cdots$	p_{lm}

Součet pravděpodobností v prvním sloupci je $p_{11} + p_{12} + \dots + p_{1m} = p_1(x_1)$, neboť je to pravděpodobnost hodnoty znaku x_1 ať y nabude kterékoliv z možných hodnot. Podobně je tomu v ostatních sloupcích. Součet pravděpodobností v prvním řádku je $p_{11} + p_{21} + \dots + p_{l1} = p_2(y_1)$, neboť je to pravděpodobnost hodnoty znaku y_1 ať x nabude kterékoliv z možných hodnot. Utvoříme-li průměr podle definice, dostaneme

$$\begin{aligned}\mathfrak{E}(x+y) = {} & (x_1 + y_1)\, p_{11} + \dots + (x_1 + y_m)\, p_{1m} + \\ & + (x_2 + y_1)\, p_{21} + \dots + (x_2 - y_m)\, p_{2m} + \\ & \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ & + (x_e + y_1)\, p_{e1} + \dots + (x_l + y_m)\, p_{lm}.\end{aligned}$$

Když roznásobíme a pak sečteme hodnoty znaku x v každém řádku, kdežto hodnoty y sečteme v každém sloupci, dostáváme

$$\begin{aligned}\mathfrak{E}(x+y) = {} & x_1 p_1(x_1) + x_2 p_1(x_2) + \dots + x_l p_1(x_l) + \\ & + y_1 p_2(y_1) + y_2 p_2(y_2) + \dots + y_m p_2(y_m).\end{aligned}$$

Z toho je vzhledem k (64) a (65) patrno, že

$$\mathfrak{E}(x+y) = \mathfrak{E}(x) + \mathfrak{E}(y),$$

což lze snadno rozšířiti na libovolný konečný počet sčítanců:

$$\mathfrak{E}(x + y + z + \dots) = \mathfrak{E}(x) + \mathfrak{E}(y) + \mathfrak{E}(z) + \dots \quad (66)$$

Obdobná věta platí pro očekávané hodnoty rozptylu. Podle definice je

$$\sigma^2(x) = \mathfrak{E}[x - \mathfrak{E}(x)]^2$$

a označíme-li $x - \mathfrak{E}(x) = \xi$, bude

$$\sigma^2(x) = \mathfrak{E}(\xi^2) = \xi_1^2 p_1(x_1) + \xi_2^2 p_1(x_2) + \dots + \xi_l^2 p_1(x_l)$$

a podobně pro $y - \mathfrak{E}(y) = \eta$

$$\sigma^2(y) = \mathfrak{E}(\eta^2) = \eta_1^2 p_2(y_1) + \eta_2^2 p_2(y_2) + \dots + \eta_m^2 p_2(y_m).$$

Pro součet obou náhodných proměnných můžeme psáti rozptyl opět pomocí hořejší tabulky

$$\sigma^2(x+y) = \mathfrak{E}(\xi+\eta)^2 = (\xi_1+\eta_1)^2\, p_{11} + \dots + (\xi_l+\eta_m)^2\, p_{lm} =$$
$$= (\xi_1^2+2\xi_1\eta_1+\eta_1^2)\, p_{11} + (\xi_1^2+2\xi_1\eta_2+\eta_2^2)\, p_{12} + \dots$$
$$\dots + (\xi_1^2+2\xi_1\eta_m+\eta_m^2)\, p_{1m} +$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$+ (\xi_l^2+2\xi_l\eta_1+\eta_1^2)\, p_{l1} + (\xi_l^2+2\xi_l\eta_2+\eta_2^2)\, p_{l2} + \dots$$
$$\dots + (\xi_l^2+2\xi_l\eta_m+\eta_m^2)\, p_{lm}.$$

Sečteme opět čtverce ξ^2 v každém řádku a čtverce η^2 v každém sloupci a dostaneme

$$\sigma^2(x+y) = \xi_1^2 p_1(x_1) + \dots + \xi_l^2 p_1(x_l) +$$
$$+ \eta_1^2 p_2(y_1) + \dots + \eta_m^2 p_2(y_m),$$

čili

$$\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y),$$

neboť součet všech součinů $\Sigma\xi_i\eta_k p_{ik}$ se rovná nule. O tom se snadno přesvědčíme, ježto jej dostaneme provedením součinu součtů

$$\sum_{i=1}^{l} \xi_i p_1(x_i) \cdot \sum_{k=1}^{m} \eta_k p_2(y_k)$$

a každý z těchto součtů je roven nule, neboť je to první moment kolem aritmetického průměru.

Také zde platí obecně

$$\sigma^2(x+y+z+\dots) = \sigma^2(x) + \sigma^2(y) + \sigma^2(z) + \dots \quad (67)$$

Podobně bychom odvodili

$$\sigma^2(x-y) = \sigma^2(x) + \sigma^2(y). \quad (67')$$

Bernoulliův problém se jeví jako nejjednodušší zvláštní případ tvoření součtů. Vzniká, když se pravděpodobnosti $p_1(x_i)$, $p_2(y_k)$, $p_3(z_j)$, ... vztahují na alternativu, takže se hodnoty každého znaku redukují na dvě, jež označíme 1, 0. Potom je

$$p_1(1) = p_2(1) = \dots = p, \quad p_1(0) = p_2(0) = \dots = q.$$

Pravděpodobnost, že v náhodném výběru bude x prvků s pozorovaným znakem c je táž, jako pravděpodobnost, že

součet jednotek bude x a tedy počet nul $r - x$. Výsledky, jež jsme našli, odpovídají právě odvozeným větám, neboť bylo $\mathfrak{E}(x) = rp$, $\sigma^2(x) = rpq$ pro rozdělení četností $P_r(x)$.

Pro očekávané hodnoty platí ještě další věty, které stačí uvésti:

$\alpha)$ $$\mathfrak{E}(a) = a,$$

kde a je konstanta. Z toho důvodu také

$\beta)$ $$\mathfrak{E}[\mathfrak{E}(x)] = \mathfrak{E}(x),$$

$\gamma)$ $$\mathfrak{E}(ax) = a\,\mathfrak{E}(x).$$

Očekávaná hodnota součinu dvou náhodných proměnných na sobě nezávislých se rovná součinu jejich očekávaných hodnot.

$\delta)$ $$\mathfrak{E}(xy) = \mathfrak{E}(x)\,\mathfrak{E}(y).$$

Dvě náhodné proměnné jsou na sobě nezávislé, zůstává-li rozdělení četností jedné proměnné stále totéž, ať druhá proměnná nabývá kterékoliv hodnoty. Říká se také, že jsou stochasticky nezávislé.

Platí dále analogická věta jako (5) mezi obecným druhým momentem a druhým momentem kolem aritmetického průměru

$\varepsilon)$ $$\mathfrak{E}(x^2) = \mathfrak{E}(\xi^2) + [\mathfrak{E}(x)]^2.$$

(7,4) Zákon velkých čísel. Můžeme nyní odvoditi podle Misesa další obecnou větu, která zahrnuje jako zvláštní případy teorém Bernoulliův i Poissonovo zobecnění tvoříc součást vět vyjadřujících princip velkých čísel.

Odvodili jsme si očekávanou hodnotu a rozptyl rozdělení pravděpodobnosti vzniklého tvořením součtů náhodných proměnných. Hledejme nyní tyto parametry nikoliv pro součet náhodných proměnných, nýbrž pro jejich průměr. Vyjdeme od r základních souborů a vezmeme z každého z nich jeden prvek; budeme na nich sledovati (pro jednoduchost) znak alternativní, který bude vyznačen 1 a 0.

Celkový součet hodnot znaků bude tedy součtem jednotek na př. x. Budeme tvořit průměry tím, že součty x dělíme počtem prvků r, tedy $f = \frac{x}{r}$. Přechod od původních základních souborů k novému se znakem f označujeme jako tvoření průměrů.

Hledáme pravděpodobnost $P_r(f)$, že z r prvků bude x prvků s pozorovaným znakem, takže dostaneme z nich průměr f. Tato pravděpodobnost průměru souvisí vztahem

$$P_r(f) = P_r(x) = P_r(rf)$$

s pravděpodobností $P_r(x)$ součtu x, jak jsme již konstatovali (str. 98).

Průměr rozdělení pravděpodobností $P_r(f)$ je

$$\mathfrak{E}(f) = \sum_f f\, P_r(f) = \sum_x \frac{x}{r} P_r(x) = \frac{\mathfrak{E}(x)}{r}. \tag{68}$$

Rozptyl

$$\sigma^2(f) = \sum_f (f - \mathfrak{E}(f))^2 P_r(f) = \sum_x \left(\frac{x - \mathfrak{E}(x)}{r}\right)^2 P_r(x) = \frac{\sigma^2(x)}{r^2}. \tag{69}$$

Značí tedy přechod od $P_r(x)$ ku $P_r(f)$ sesunutí úseček (obr. 12) v poměru r: 1. Poněvadž pro r základních souborů jsou očekávané hodnoty $\mathfrak{E}(x_1), \ldots, \mathfrak{E}(x_r)$ a tedy podle (66) $\mathfrak{E}(x) = \mathfrak{E}(x_1) + \mathfrak{E}(x_2) + \ldots + \mathfrak{E}(x_r)$, bude vzhledem k (68)

$$\mathfrak{E}(f) = \frac{\mathfrak{E}(x_1) + \mathfrak{E}(x_2) + \ldots + \mathfrak{E}(x_r)}{r}.$$

Podobně můžeme psáti vzhledem k (67) a (69) rozptyl

$$\sigma^2(f) = \frac{\sigma^2(x_1) + \sigma^2(x_2) + \ldots + \sigma^2(x_r)}{r^2}.$$

Zavedeme-li předpoklad, že rozptyly $\sigma^2(x_i)$ těch jednotlivých rozdělení četností mají horní hranici σ^2, že tedy $\sigma^2(x_i) \leqq \sigma^2$

pro $i = 1, 2, \ldots, r$ pak z poslední rovnice plyne, že

$$\sigma^2(f) \leqq \frac{\sigma^2}{r}$$

čili $\lim_{r\to\infty} \sigma^2(f) = 0$.

Rozptyl rozdělení pravděpodobností $P_r(f)$ spěje s rostoucím r k nule právě jako v případě Bernoulliově.

Pravděpodobnost, že f bude v mezích $\pm z_0$ kolem průměru, bude vymezena zase nerovninou

$$P_r(\mathfrak{E}(f) - z_0, \mathfrak{E}(f) + z_0) \geqq 1 - \frac{\sigma^2}{rz_0^2}$$

čili

$$\lim_{r\to\infty} P_r(\mathfrak{E}(f) - z_0, \mathfrak{E}(f) + z_0) = 1.$$

Můžeme tedy vyslovit větu:

Pravděpodobnost, že průměr r veličin, z nichž každá podléhá nějakému libovolnému rozdělení pravděpodobností, leží v libovolně malém intervalu u své očekávané hodnoty, je libovolně blízka 1, když r je dosti velké. Předpokladem je, že rozptyly $\sigma^2(x_i)$ jednotlivých rozdělení mají určitou horní hranici, nebo jejich součet roste slaběji než r^2. Lze také říci stručněji: Při velkém r je skoro jisto, že průměr čísel, která podléhají nějakým r rozdělením, bude přibližně roven své očekávané hodnotě.

(8,1) Odhad parametrů základního souboru podle příslušných charakteristik výběrových.

Dosud jsme se zabývali hlavně otázkou, co můžeme říci o relativní četnosti f pozorovaného znaku v náhodných výběrech, známe-li jeho relativní četnost p v příslušném základním souboru, z něhož byly vzaty. Odvodili jsme velmi užitečné věty o rozptylu alternativního znaku v náhodných výběrech.

Při statistické praksi však je častěji třeba usuzování směrem obráceným. Ze znalosti charakteristiky v jednom

nebo několika pozorovaných výběrech máme odhadnouti neznámou hodnotu příslušného parametru v základním souboru. K tomu cíli hledáme odpověď hlavně na tyto čtyři typy otázek:

1. Jaká je pravděpodobnost určité hodnoty neznámého parametru?

2. Jaký je tudíž rozptyl jeho hodnot?

3. Kterou hodnotu máme podle pozorování určitého náhodného výběru považovati za nejbližší a tedy nejlepší hodnotu neznámého parametru?

4. Lze považovati dva nebo několik souborů za náhodné výběry z téhož základního souboru?

Statistikovým úkolem tedy je především, udati na základě pozorovaného výběru meze, v nichž je neznámý parametr základního souboru, čili stanoviti jeho rozptyl a najíti, kterou hodnotu lze pro tento parametr pokládati za nejlepší.

(8,2) Meze základní relativní četnosti. Poněvadž se v tomto oddílu zabýváme jen znakem alternativním, budeme řešiti naznačené úkoly pro relativní četnost f a ji odpovídající parametr p.

Řešení nám zase usnadní Laplaceův integrál, který udává pravděpodobnost $\alpha(t)$, že odchylka četnosti x od průměru $\bar{x} = rp$ bude v mezích $\pm t\,\sigma(x)$, čili s pravděpodobností

$$\alpha(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{1}{2}\tau^2} \mathrm{d}\tau \tag{70}$$

platí nerovnosti

$$-t\,\sigma(x) \leqq x - rp \leqq +t\,\sigma(x). \tag{71}$$

Znamená to, že v souboru, který má za prvky všechny kombinace r prvků z celkového počtu N, a má tedy rozsah $\binom{N}{r}$, existuje zcela určitá relativní četnost takových kombi-

nací, v nichž počet prvků s pozorovaným znakem se neodchyluje od rp více než o $t\,\sigma(x)$ dolů nebo nahoru. Určitými nerovnostmi (71) je v daném souboru stanovena pravděpodobnost (70); také obráceně, předepíšeme-li si určitou pravděpodobnost, (70) plynou z ní přímo určité nerovnosti (71); Tyto nerovnosti můžeme psáti také v jiném tvaru, přičteme-li na každé straně rp

$$rp - t\,\sigma(x) \leqq x \leqq rp + t\,\sigma(x)$$

nebo

$$p - \frac{t\,\sigma(x)}{r} \leqq \frac{x}{r} \leqq p + \frac{t\,\sigma(x)}{r} \tag{72}$$

$$p - t\sqrt{\frac{pq}{r}} \leqq f \leqq p + t\sqrt{\frac{pq}{r}}.$$

Tím je tedy relativní četnost $\frac{x}{r} = f$ sevřena do určitých mezí při daném p, t, r, N, neboť směrodatná odchylka $\sigma(x)$ je buď $\sqrt{rpq}$, nebo $\sqrt{rpq\left(1 - \frac{r}{N}\right)}$, nevrací-li se při provádění výběru prvky do základního souboru konečného rozsahu N.

Jedná se nám nyní o to, abychom odvodili přípustné meze, v nichž musí býti p při určitém, daném $\frac{x}{r} = f$. Dolní hranice (72) je $-\frac{t\,\sigma(x)}{r} = f - p$ a horní hranice $+\frac{t\,\sigma(x)}{r} = f - p$. Jejich čtverec je týž, a dosadíme-li v něm za $\sigma(x)$ druhý obecnější výraz, máme

$$t^2 p\,(1-p)\left(\frac{1}{r} - \frac{1}{N}\right) = (f-p)^2.$$

To je rovnice druhého stupně pro p a jejím řešením dostáváme dva kořeny

$$p = f + \left\{t^2\left(\frac{1}{2} - f\right)\left(\frac{1}{r} - \frac{1}{N}\right) \pm t \cdot \right.$$
$$\left. \cdot \sqrt{f(1-f)\left(\frac{1}{r} - \frac{1}{N}\right) + \frac{t^2}{4}\left(\frac{1}{r} - \frac{1}{N}\right)^2}\right\} : \left\{1 + t^2\left(\frac{1}{r} - \frac{1}{N}\right)\right\}, \tag{73}$$

které tvoří horní a dolní mez pro p. Tento výsledek může býti zjednodušen především tím, že klademe výraz v děliteli přibližně roven jedné, neboť vzhledem k $t = \frac{\xi}{\sqrt{rpq\left(1 - \frac{r}{N}\right)}}$ bude

$$1 + t^2\left(\frac{1}{r} - \frac{1}{N}\right) = 1 + \frac{\xi^2 \frac{1}{r}\left(1 - \frac{r}{N}\right)}{rpq\left(1 - \frac{r}{N}\right)} = 1 + \frac{\xi^2}{r^2 pq}$$

a veličiny řádu $\frac{\xi^2}{r^2 pq}$ jsme při odvozování křivky Gaussovy zanedbávali, takže také zde můžeme zůstati v obdobných mezích přibližnosti.

Dostali jsme tak pro relativní četnost v základním souboru nerovnosti, jimiž je sevřena při známé relativní četnosti výběrové f

$$f + t^2\left(\frac{1}{2} - f\right)\left(\frac{1}{r} + \frac{1}{N}\right) -$$
$$- t\varrho \leqq p \leqq f + t^2\left(\frac{1}{2} - f\right)\left(\frac{1}{r} - \frac{1}{N}\right) + t\varrho, \tag{74}$$

kde

$$\varrho = \sqrt{f(1-f)\left(\frac{1}{r} - \frac{1}{N}\right) + \frac{t^2}{4}\left(\frac{1}{r} - \frac{1}{N}\right)^2}.$$

Je-li rozsah r tak velký, že stačí přihlížeti k veličinám řádu $\frac{1}{\sqrt{r}}$ a zanedbati veličiny řádu $\frac{1}{r}$, dostaneme přibližné ne-

rovnosti

$$f - t\sqrt{f(1-f)\left(\frac{1}{r} - \frac{1}{N}\right)} \leqq p \leqq f + t\sqrt{f(1-f)\left(\frac{1}{r} - \frac{1}{N}\right)} \tag{75}$$

a pro základní soubor nekonečného rozsahu $N = \infty$ čili pro případ výběru s vracením prvků

$$f - t\sqrt{\frac{f(1-f)}{r}} \leqq p \leqq f + t\sqrt{\frac{f(1-f)}{r}}. \tag{76}$$

Je zřejmo, že nerovnosti (76) jsou inversí nerovností (72), neboť p a f si vyměnily místo. Nerovnosti (76) tedy udávají hranice, v nichž je sevřena pravděpodobnost znaku p při dané relativní četnosti f a určitě zvoleném t s pravděpodobností $\alpha(t)$. Velký praktický význam této inverse je v tom, že dostáváme i při neznámém p dobré přiblížení pro $\alpha(t)$ z tabulky Laplaceova integrálu, nahradíme-li p ve výrazech pro směrodatnou odchylku hodnotou f, kterou jsme stanovili z výběru. Použijeme pak zase věty Cournotovy, abychom přešli od matematických výsledků k závěrům o skutečnosti. Nejprve si stanovíme určitou nejmenší hranici pro pravděpodobnosti, na něž ještě chceme bráti zřetel. Potom považujeme hodnoty znaku nebo odchylky, jejichž celková pravděpodobnost je menší, za „velmi zřídka se vyskytující" nebo „prakticky se nevyskytující". Tyto nejmenší hranice se v literatuře nazývají také „fiduciální meze", nebo „interval konfidence". Rozhodneme se na příklad, že nebudeme přihlížeti k pravděpodobnostem $0{,}0027 = 1 - \alpha(t)$; tato hranice odpovídá hodnotě $t = 3$. Tím říkáme, že odchylky od průměru větší než $\pm 3\sigma_x$ pozorujeme v náhodném výběru z dobře promíchaného základního souboru „velmi zřídka" nebo „prakticky nikdy". Potom určíme pomocí této hodnoty $t = 3$ meze pro p v nerovnostech (74), (75) nebo (76). Konečně pak vyvodíme závěr, který je obrácením Cournotovy formulace zákona velkých čísel a odpovídá na otázku,

v jakých mezích je neznámý parametr p takto: Přihodí se „velmi zřídka" nebo „prakticky nikdy", aby pravděpodobnost pozorovaného znaku byla vně určených hranic (pro $t = 3$), byl-li vzat náhodný výběr rozsahu r ze základního souboru dobře promíchaného. Předpokladem je, že r je tak velké, že užití Laplaceovy formule je přípustné.

(8,3) Přibližná hodnota parametru p. Nyní máme dáti odpověď na druhou otázku: Kterou přibližnou hodnotu máme nejlépe přijmouti pro parametr p. Obyčejně se považuje za nejlepší přibližnou hodnotu pro p relativní četnost f nejčastěji pozorovaná, t. j. ta, která má v základním souboru největší relativní četnost, nebo průměr všech hodnot f, které se vyskytují. Obě cesty, o nichž se blíže zmíníme až v druhém díle, vedou zde k témuž výsledku, že za nejlepší přibližnou hodnotu parametru p bereme pozorovanou relativní četnost f. Trochu jiný výsledek dostaneme, když vyjdeme od nerovností (74), neboť tam vidíme, že se odchylky nepočítají od f nýbrž od

$$f + t^2\left(\frac{1}{2} - f\right)\left(\frac{1}{r} - \frac{1}{N}\right),$$

což nás může vésti k hodnotě opravené druhým členem, který vymizí pro $f = \frac{1}{2}$ a je tím větší, čím je f vzdálenější od $\frac{1}{2}$ a čím je větší t. Tato oprava posunuje vždy přibližnou hodnotu f blíže k $\frac{1}{2}$. Můžeme si uvésti pro $N = \infty$ několik čísel pro ilustraci. Pro $r = 100$, $t = \sqrt{10}$ dostáváme při pozorovaném f

0,10 0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90

opravenou přibližnou hodnotu pro parametr p

0,14 0,23 0,32 0,41 0,50 0,59 0,68 0,77 0,86.

Třebaže nemůžeme stanoviti jednoznačně přibližnou hodnotu pro p, poněvadž stojí v cestě obtíže vyplývající z povahy problému, přece je opravdovým úspěchem matema-

tické teorie, že můžeme dosíci za určitých podmínek velmi cenných odhadů parametru tím, že lze zkoumati a odhadnouti rozptyl resp. směrodatnou odchylku hodnot, z nichž jsme jednu zjistili náhodným výběrem.

Očekávaná hodnota $\mathfrak{E}(f)$ relativní četnosti $f = \frac{x}{r}$, která je průměrem jejích hodnot, zjištěných ve všech možných výběrech rozsahu r je rovna příslušnému parametru p v základním souboru. Jako přibližnou jeho hodnotu dostáváme relativní četnost $\frac{x}{r}$ z pozorovaného náhodného výběru, která je mu tím bližší, čím je r větší. Potřebujeme tedy vyjádřit očekávanou hodnotu rozptylu $\mathfrak{E}(\sigma^2)$ pomocí pozorovaných hodnot přibližných. Víme již (str. 69), že očekávaná hodnota rozptylu relativních četností je $\frac{pq}{r}$, jakožto hodnota rozptylu v základním souboru. Známe však pro p a q jen přibližné hodnoty f a $1 - f$. Nemůžeme vzíti za přibližnou hodnotu rozptylu jednoduše $\frac{1}{r} f(1-f)$, která by vyplývala, kdybychom kladli za p přibližnou hodnotu f. O tom se přesvědčíme, když si vypočítáme, jaká by byla očekávaná hodnota součinu $\frac{x}{r}\left(1 - \frac{x}{r}\right)$ daného výsledkem pozorování. Stanovíme tedy očekávanou hodnotu výrazu $\frac{x}{r} - \frac{x^2}{r^2} = \frac{x}{r}\left(1 - \frac{x}{r}\right)$. Očekávaná hodnota $\frac{x}{r}$ je p; očekávaná hodnota $\frac{x^2}{r^2}$ je podle věty (γ) rovna $\frac{pq}{r} + p^2$, neboť druhý moment kolem průměru je $\frac{pq}{r}$ a čtverec očekávaného průměru, který je totožný s průměrem v základním souboru je p^2.

Bude tudíž celková očekávaná hodnota uvažovaného součinu podle (66)

$$p - \frac{pq}{r} - p^2 = p(1-p) - \frac{pq}{r} = pq\left(1 - \frac{1}{r}\right)$$

a je odlišná od součinu pq. Kdybychom přijali $\frac{x}{r}\left(1 - \frac{x}{r}\right)$ za přibližnou hodnotu očekávané hodnoty pq, dopouštěli bychom se tedy jednak chyby systematické, jež se jeví v součiniteli $\frac{r-1}{r}$, jednak druhé chyby v tom, že existuje odchylka mezi zvláštní pozorovanou hodnotou $\frac{x}{r}$ náhodné proměnné a její očekávanou hodnotou p. Systematickou chybu můžeme opravit tím, že vezmeme za přibližnou hodnotu $\frac{r}{r-1}\frac{x}{r}\left(1 - \frac{x}{r}\right)$, neboť její očekávaná hodnota je pak právě pq. Z toho tedy vyplývá, že

$$\frac{1}{r}\frac{x}{r}\left(1 - \frac{x}{r}\right) = \frac{r-1}{r}\frac{pq}{r}. \tag{77}$$

Další otázkou, kdy lze považovati dva nebo více souborů za náhodné výběry z téhož základního souboru, budeme se zabývati v druhém díle. Zde se omezíme jen na konvenci, která se ujala dnes ve statistice. Dostaneme-li pro jednu charakteristiku, v našem případě pro relativní četnost ze dvou různých výběrů hodnoty f_1 a f_2, považujeme souhlas mezi nimi za dobrý, když rozdíl $|f_1 - f_2|$ je menší než směrodatná odchylka této diference podle (67′) tedy $\sqrt{\sigma_{f_1}^2 + \sigma_{f_1}^2}$, a za uspokojivý, je-li menší než dvojnásobek, někdy i trojnásobek jeho směrodatné odchylky.

Přesahuje-li rozdíl trojnásobek směrodatné odchylky, nepovažuje se souhlas za uspokojivý a vzniká domněnka, že lze najíti vysvětlení této odchylky zvláštní příčinou, nikoliv náhodným výběrem.

(8,4) Pearsonovo kriterium χ^2. Podle výsledků, jež jsme dosud odvodili, můžeme stanovit meze, v nichž je relativní četnost v základním souboru p sevřena, zvolíme-li si za přípustný interval odchylek délku $\pm 3\sigma_x$. Tak pro tabulku 1. našeho příkladu (str. 29) máme pro hodnotu třídního znaku $x_i = 75$ relativní četnost $f_i = 0{,}141$, takže směrodatná odchylka $\sigma_f = 0{,}021$ a tudíž meze jsou $f_i \pm 0{,}063$. Tak si můžeme vypočítati meze pro parametr p_i každé třídy z pozorovaného rozdělení četností.

Klademe si však dále otázku, jak bychom vystihli, do jaké míry se liší rozdělení pozorovaného souboru jako celek od základního souboru, nikoliv jak se liší jednotlivé četnosti od příslušných parametrů. V odpověď na to sestrojil K. Pearson t. zv. kriterium χ^2.

V základním souboru jsou statistické pravděpodobnosti hodnot znaku kvantitativního resp. třídních hodnot znaku $p_1, p_2, \ldots, p_l$. Výběr rozsahu r, který by měl tytéž relativní četnosti, by vykazoval třídní četnosti $\nu_1 = rp_1, \ldots, \nu_l = rp_l$. Tyto četnosti porovnáváme s pozorovanými rf_i tak, že tvoříme jejich rozdíly; čtverce rozdílů pak vyjádříme v poměru k teoretickým četnostem ν_i a sečteme. Tak dostaneme výraz

$$\chi^2 = \sum_{i=1}^{l} \frac{(rf_i - rp_i)^2}{rp_i}.$$

Všechny čtverce rozdílů se sčítají a je zřejmo, že čím jsou rozdíly obojích četností větší, tím je větší χ^2; jsou-li obě rozdělení shodná, je $\chi^2 = 0$. Uvedený výraz můžeme také psáti v tvaru

$$\chi^2 = \sum_{i=1}^{l} \frac{r(f_i - p_i)^2}{p_i} = r\left(\sum_{i=1}^{l} \frac{f_i^2}{p_i} - 1\right)$$

neboť

$$\sum_{i=1}^{l} r\left(\frac{f_i^2}{p_i} - 2f_i + p_i\right) = r\sum_{i=1}^{l} \frac{f_i^2}{p_i} - 2r + r,$$

vzhledem k tomu, že

$$\sum_{i=1}^{l} f_i = \sum_{i=1}^{l} p_i = 1.$$

Vidíme, že je to charakteristika, vztahující se k určitému výběru rozsahu r, která nám podává zhuštěnou informaci o tom, jak se tento výběr v celku liší svým rozdělením četností od očekávaného. Pro každý výběr bychom dostali pravděpodobně jinou hodnotu, takže ze všech $\binom{N}{r}$ hodnot bude utvořeno rozdělení četností této charakteristiky. K tomuto rozdělení četností se utvoří součtová křivka $F_1(\chi^2)$ integrací obdobně jako jsme dostali Laplaceův integrál (53) nebo (50), která však závisí ještě na druhé veličině $l - 1$, kde l je počet tříd.

Z ní se tedy dovíme, jaká je pravděpodobnost, že při určitém r a daných hodnotách p_i dostaneme větší hodnotu pro χ^2, než je pozorovaná. Je to tedy pravděpodobnost, s níž můžeme očekávati horší souhlas s teoretickým rozdělením, než je pozorovaný.

Přesvědčíme se, jak vystihuje v příkladu 3. str. 97 teoretické rozdělení četností pomocí dvou členů řady Poisson-Charlierovy rozdělení pozorované tím, že vypočítáme charakteristiku χ^2.

x_i	n_i	rp_i	$\lvert n_i - rp_i \rvert$	$(n_i - rp_i)^2$	$\frac{(n_i - rp_i)^2}{rp_i}$
5	133	135,2	2,2	4,84	0,04
6	55	51,2	3,8	14,44	0,28
7	23	22,5	0,5	0,25	0,01
8	7	9,6	2,6	6,76	0,70
9	2	2,8	0,8	0,64	0,23
10	2	0,7	1,3	1,69	2,41
Σ	222	222,0			3,67

Vidíme, že $\chi^2 = 3{,}67$ a podle příslušné tabulky Eldertonovy mu odpovídá pro $l - 1 = 5$ $F_1(\chi^2) = 0{,}60$, což je pravděpodobnost, že dostaneme v náhodných výběrech větší hodnoty χ^2, než je pozorovaná; bylo by to tedy přibližně v 60 případech ze 100. Takové vystižení není příliš dobré. Ovšem v tomto případě se uplatňuje příliš vliv posledních dvou málo obsazených tříd; příslušná čísla rp_i ve jmenovateli pak příliš zvyšují hodnotu χ^2, jak vidíme na poslední třídě a není to tedy jen vlivem rozdílů. Proto a také z důvodů spočívajících v odvození, jež předpokládá, že odchylky od očekávaných četností vyhovují normální křivce, se obyčejně krajní třídy málo obsazené spojují dohromady, aby četnost byla aspoň 5.

Spojíme-li tedy poslední dvě třídy, bude pak $\chi^2 = 1{,}1$ a jemu odpovídá pro $l - 1 = 4$ pravděpodobnost $F_1(\chi^2) = 0{,}78$. Z toho můžeme usuzovati, že bychom dostali přibližně v 78 případech náhodných výběrů ze sta řadu pozorovaných četností, jež dává skupinu odchylek od teoretického rozdělení tedy χ^2, jež je méně pravděpodobné než pozorované; měli bychom tedy očekávati zhruba v každém stu náhodných výběrů 78krát horší souhlas s teoretickým, než je pozorovaný, vyjádřený charakteristikou $\chi^2 = 1{,}1$.

(8,5) Příklady. 1. Roční míra úmrtnosti 60letých osob byla v nějakém rozsáhlém souboru zjištěna a uvedena v tabulce úmrtnosti $q_{60} = 0{,}0287$. Jaká je pravděpodobnost náhodné odchylky menší než $\pm z_0 = 0{,}01$ roční míry pozorované v souboru rozsahu $r = 2500$ a menší než $\pm z_0 = 0{,}005$ v souboru rozsahu $r = 10\,000$.

Tuto pravděpodobnost udává Laplaceův integrál, jehož horní mez určíme ze vztahu $\gamma_0 = z_0 \sqrt{\frac{r}{2pq}} = z_0 \sqrt{\frac{r}{2f(1-f)}}$, kde bude $z_0 = 0{,}01$, $r = 2500$, $f = 0{,}0287$, $1 - f = 0{,}9713$. Tak dostáváme $\gamma_0 = 1{,}339$ a tedy $\Phi(\gamma_0) = 0{,}997$ a stejnou hodnotu máme v druhém případě.

2. Mezi 1 359 671 narozenými chlapci bylo 58 744 mrtvě narozených a mezi 1 285 086 děvčaty 44 224 mrtvě narozených. Vypočítejte podle pohlaví procento mrtvě narozených, které je

průměrem za několik let, z nichž jsou tato čísla úhrnem udána, pak rozdíl obou procent a jejich rozptyl. Jaké jsou závěry z těchto čísel?

Mrtvě narozených mezi chlapci bylo 4,320%, rozptyl je 0,017%
„ „ „ děvčaty „ 3,441%, „ „ 0,016%
Rozdíl těchto procent je 0,879%, „ „ 0,023%.

Mezi chlapci a děvčaty je skutečně rozdíl v embryonální úmrtnosti, což se potvrzuje také rozsáhlejším jiným statistickým materiálem.

3. Pravidelné zkoušky obsahu popela v uhlí (Pearson) daly průměr $\bar{x} = 16{,}99\%$ a směrodatnou odchylku $\sigma_x = 2{,}66\%$. Kolik vagonů z dodávky 250 vagonů bude míti obsah popela mezi 12 a 20%? Meze zkoumaného intervalu, vyjádřené ve směrodatné proměnné budou $t_1 = \frac{12 - 16{,}99}{2{,}66} = -1{,}88$ dolní a $t_2 = \frac{20 - 16{,}99}{2{,}66} = +1{,}13$ horní. Plocha křivky Laplace-Gaussovy mezi pořadnicemi příslušnými těmto mezím je $47{,}00 + 37{,}18 = 84{,}18$ procent. Bylo by podle toho $250 \times 0{,}8418 = 210$ vagonů s udaným obsahem popela; skutečně jich bylo 211.

4. Znázorníme-li si graficky, jaké procento impregnovaných dřevěných sloupů, zasazených do suché půdy uhnívá v jednotlivých letech, dostaneme přibližně normální křivku. Sestrojíme-li si k ní součtovou křivku, bude to křivka života těchto sloupů. Srovnej s obr. 15. Pomocí ní snadno vidíme poločas — t. j. dobu, kdy uhnila polovička sloupů, vedeme-li rovnoběžnou úsečku s osou roků, ve výši 50%, je potom $\tilde{x} = 12$ roků; podobně čas čtvrtinový $\tilde{x}_1 = 10{,}5$ jakož i minimální život, t. j. čas, kdy uhnije první sloup z řady současně postavených $x_1 = 6$ a maximální $x_l = 18$, kdy uhnily všechny.

5. Při třídění cihel bylo zjištěno za celý rok 70% cihel I. jakosti. Při posledním vyvážení z pece bylo v 10 000 kusech jen 68% I. jakosti. Lze považovati tuto odchylku za nahodilou? Relativní četnost $f = 0{,}7$, $1 - f = 0{,}3$ tedy

$$\sigma_f = \sqrt{0{,}7 \times 0{,}3 : 10\,000} = 0{,}0046.$$

Vzhledem k tomu, že odchylka 2% překračuje značně interval $\pm 3\sigma_f = 1{,}38\%$, který za normálních podmínek křivky Laplace-

Gaussovy je překročen s pravděpodobností 0,0027, nepovažuje se v praksi za odchylku nahodilou.

6. Ve sklárně se zjistí, že automat na výrobu lahví dal při přejímací zkoušce 2% vadných lahví ze 4000 kusů udělaných při zkoušce. Jaké bude asi mezní procento výmětů při plynulé výrobě? Vezmeme tedy za přibližnou hodnotu očekávané hodnoty $f = 0{,}98$, $1 - f = 0{,}02$, takže

$$\sigma_f = \sqrt{0{,}02 \times 0{,}98 : 4000} = 0{,}223 \text{ procent},$$

takže mez pro výměty je pravděpodobně $2 + 3 \times 0{,}223 = 2{,}7$ procent.

7. Určité křížení hrachu dalo 5321 žlutých a 1804 zelená zrnka. Podle hypotesy Mendelovy je očekávaný počet zelených zrnek 25%. Lze považovati tuto odchylku od očekávané hodnoty za vzniklou jen náhodným výběrem?

Odchylka pozorovaného výsledku od očekávaného je $\xi = 23$. Směrodatná odchylka $\sigma_f = \sqrt{0{,}25 \times 0{,}75 \times 7125} = 36{,}6$. Poněvadž odchylka ξ je jen asi $0{,}6\sigma_f$, mohla vzniknouti zcela dobře jen náhodným výběrem.

8. Za víceleté období se objevil ve statistice dětských sebevražd roční průměr $\overline{x} = 1{,}96$. Můžeme uvésti jako příklad použití Poissonovy exponentiely tento jev a nepotřebujeme znáti explicitně r a p. Tak dostaneme pravděpodobnost, že se v jednom roce nevyskytne ani jedna sebevražda, pak že se vyskytne v jednom roce jedna, dvě, atd. Z rovnice (54) dostáváme $\psi(0) = 0{,}141$, $\psi(1) = 0{,}276$, $\psi(2) = 0{,}271$ $\psi(3) = 0{,}177$, $\psi(4) = 0{,}087$. Součet těchto pravděpodobností je 0,952, takže na všechny ostatní dohromady zbývá 4,8%, což je pravděpodobnost více než čtyř případů dětských sebevražd v roce. Největší pravděpodobnosti mají případy $x = 1$ a $x = 2$, mezi nimiž leží průměr, a to blíže k $x = 2$ hlavně v důsledku toho, že $\psi(3)$ je větší než $\psi(0)$.

(9,1) Lexisova teorie. Všimli jsme si, že pro teorém Bernoulliův a teorii s ním související až na zobecnění Poissonovo je podstatným znakem, že pravděpodobnost p, která je podkladem relativních četností získaných pozorováním, je konstantní. Pozorované statistické soubory bývají složeny z prvků mnohotvárnějších a složitějších než odpovídá schematu Bernoulliovu. Zakladatelé matematické statistiky, i Laplace, považovali totožnost pozorované statistické

řady s řadou Bernoulliovou za samozřejmou. Teprve Lexis ukázal nepostačitelnost dosavadních úvah a podal jasnější pohled na povahu statistických řad. Používání směrodatné odchylky (40) pro rozbor pozorovaných řad dává příliš hrubé výsledky, které jsou tím vzdálenější od skutečnosti, čím její podklad se více liší od podkladu typické binomické řady. Kolísání numerických hodnot pozorovaného znaku na prvcích souboru se neřídí jednoduchými zákony jako schema Bernoulliovo, působí-li na statisticky studovaný jev rušivě vnější vlivy, a proto potřebujeme míru k hodnocení zjištěných rozdílů. Tuto míru dává Lexisova teorie řad. Lexis a současně Dormoy, formuloval otázku, jak určit míry podobnosti nebo rozdílu mezi strukturou statistické řady pozorované a příslušné binomické.

Tomuto určení slouží srovnávání rozptylů resp. směrodatných odchylek řad, s nimiž se potkává statistická prakse; k němu užívá Lexisova teorie tří typů statistických řad jako norem. Metoda rozboru pak spočívá v tom, že pozorovaný soubor se rozloží na částečné soubory, v nichž by mohly býti zkoumány změny relativní četnosti znaku. Hledisko pro odvození těchto částečných souborů není dáno jen všeobecnými zásadami, nýbrž uplatněním statistikových zkušeností a znalostí vědního oboru, do něhož spadá studium pozorovaného souboru, jakož i podrobné znalosti původního materiálu a jeho pramenů. K objevení a vysvětlení podstatných změn lze pak proniknouti především statistickým uměním, které pomáhá zvoliti vhodný vědecký postup. Tyto všeobecné úvahy dále objasníme na pozorovaném materiálu. Nyní se seznámíme s uvedenými třemi typy řad, odpovídajícími jednoduchým schematům náhodných her. Srovnáváním s nimi je osvětlována náhodná stránka ve statistickém dění.

1. S prvním typem řad jsme se již seznámili. Je představován řadou, jejíž základní pravděpodobnost p výskytu pozorovaného znaku je konstantní a nazývá se řadou Bernoulliovou. Její očekávaná hodnota průměru podle (38)

je $\mathfrak{E}(x) = rp$ a očekávaná hodnota směrodatné odchylky (40) teoretického rozdělení četností byla odvozena ve výrazu $\sigma(x) = (rpq)^{\frac{1}{2}}$; očekávaná hodnota směrodatné odchylky příslušného rozdělení relativních četností je dána výrazem $\left(\frac{pq}{r}\right)^{\frac{1}{2}}$ a očekávaná hodnota průměru je p.

Uvažme nyní, že neznáme hodnotu pravděpodobnosti p, nýbrž jen pozorované hodnoty relativních četností $f_i = \frac{x_i}{r}$ z n výběrů rozsahu r prvků.

Musíme pak vzíti za přibližnou hodnotu parametru p zlomek, který je průměrem pozorovaných hodnot f_i

$$\bar{f} = \frac{1}{n}(f_1 + f_2 + \cdots + f_n) = \frac{1}{rn}(x_1 + x_2 + \cdots + x_n)$$

a za této hypotésy je očekávaná hodnota $\mathfrak{E}(\bar{f}) = p$ a také očekávaná hodnota každé jednotlivé relativní četnosti $\mathfrak{E}(f_i) = p$. Dále očekávaná hodnota

$$\mathfrak{E}(f_i - p)^2 = \frac{pq}{r}, \tag{78}$$

neboť je to průměr čtverců odchylek relativních četností z výběru rozsahu r od jejich průměru, čili rozptyl vyjádřený pomocí hodnot základního souboru.

Dále je

$$\mathfrak{E}(\bar{f} - p)^2 = \frac{1}{n^2}\mathfrak{E}\left[\sum_1^n (f_i - p)\right]^2 = \frac{pq}{r \,.\, n} \tag{79}$$

vzhledem k tomu, že

$$\bar{f} - p = \frac{1}{n}[(f_1 - p) + (f_2 - p) + \cdots + (f_n - p)] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(f_i - p) \tag{80}$$

a očekávané hodnoty součinů $\mathfrak{E}(f_i - p)(f_j - p)$, kde $i \neq j$

jsou rovny nule, ježto očekávaná hodnota každého činitele se rovná nule. Zbývá tedy n čtverců a očekávaná hodnota každého z nich je podle (78) rovna $\frac{pq}{r}$. Z toho je patrno, že každá relativní četnost f_i je přibližnou hodnotou s rozptylem $\frac{pq}{r}$ a jejich průměr $\bar{f}$ je přibližnou hodnotou, která je bližší ve smyslu teorie pravděpodobnosti s menším rozptylem $\frac{pq}{r} \cdot \frac{1}{n}$.

Statistická řada pozorovaných relativních četností má tedy projevovat rozptyl $\frac{pq}{r}$ kolem hodnoty základního souboru p.

Známe však jen přibližnou hodnotu $\bar{f}$ parametru p, takže musíme zkoumati rozptyl pozorovaných relativních četností f_i kolem jejich průměru $\bar{f}$; při tom musíme míti na paměti, že budou hráti svoji roli uvedené již dvě odchylky (str. 115).

Abychom stanovili očekávanou hodnotu tohoto rozptylu, vypočítáme si nejprve očekávanou hodnotu čtverců odchylek od přibližného průměru $\bar{f}$, takže

$$\mathfrak{E}(f_i - \bar{f})^2 = \mathfrak{E}[(f_i - p) - (\bar{f} - p)]^2 =$$

$$= \frac{pq}{r} + \frac{pq}{rn} - 2\mathfrak{E}(f_i - p)(\bar{f} - p),$$

neboť očekávané hodnoty čtverců známe podle rovnic (78) a (79) a očekávanou hodnotu posledního součinu stanovíme za předpokladu, že výběry jsou na sobě nezávislé, takže dosadíme-li tam z rovnice (80) vidíme, že $n - 1$ očekávaných hodnot součinů, kde $i \neq j$, je rovno nule a zůstává jediný $\frac{1}{n}(f_i - p)^2$, jehož očekávaná hodnota je $\frac{pq}{rn}$.

Je tedy

$$\mathfrak{E}(f_i - \bar{f})^2 = \frac{pq}{r}\left(1 - \frac{1}{n}\right).$$

Poněvadž rozptyl řady empirických hodnot kolem jejich průměru je $\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})^2$, bude také

$$\mathfrak{E}\,\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})^2 = \frac{pq}{r}\left(1 - \frac{1}{n}\right). \qquad (81)$$

Vzhledem k (78) je zřejmě $\frac{pq}{r}$ očekávanou hodnotou průměru čtverců odchylek pozorovaných relativních četností od p.

$$\mathfrak{E}\,\frac{1}{n}\sum_{i=1}^{n}(f_i - p)^2 = \frac{pq}{r}.$$

To je rozptyl řady Bernoulliovy, který budeme označovat σ_B^2. Pro přibližnou hodnotu σ_B^2 tedy stačí vzhledem k rovnici (81) bráti výraz

$$\frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2,$$

jehož očekávaná hodnota je právě σ_B^2.

Máme tudíž dva výrazy pro přibližnou hodnotu rozptylu σ_B^2. Jednak tvoříme t. zv. hodnotu počítanou $\frac{\bar{f}(1-\bar{f})}{r} = \sigma_{\bar{f}}^2$, jednak t. zv. hodnotu měřenou σ_f^2. Za předpokladu stálého složení v základním souboru a nezávislosti prvků náhodně vybíraných, mohou se tyto dvě hodnoty málo lišit, takže jejich podíl (nebo jeho odmocnina)

$$Q^2 = \frac{\sigma_f^2}{\sigma_{\bar{f}}^2}$$

musí býti blízký jednotce. Říkáme pak, že statistická řada

má normální rozptyl, je-li $Q = 1$. Pozorování je pak representováno schematem Bernoulliovým, jestliže seskupení relativních četností f_i kolem jejich průměru $\bar{f}$ odpovídá binomickému rozdělení.

Můžeme jej také vyjádřit podrobněji

$$Q^2 = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2 : \frac{\bar{f}(1-\bar{f})}{r} = \sigma_f^2 : \frac{n-1}{n}\,\frac{\bar{f}(1-\bar{f})}{r}, \quad (82)$$

kde

$$\sigma_f^2 = \frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{n}.$$

2. Druhý typ si přiblížíme představou n zalidněných okresů, v nichž pozorujeme úmrtnost x-letých (na př. 30letých) mužů; tato pravděpodobnost je v každém okresu jiná, ale konstantní. Případ si znázorníme modelem, sestrojeným z n osudí $O_1, O_2, \ldots, O_n$. Stálá pravděpodobnost vytažení černé kuličky z osudí O_1 budiž $p_1, \ldots$, a z O_n budiž p_n.

	$\overbrace{\quad\quad}^{r}$
O_1	$p_1\ p_1 \cdots p_1$
O_2	$p_2\ p_2 \cdots p_2$
$\vdots$	$\vdots\ \vdots\ \ \vdots$
O_n	$p_n\ p_n \cdots p_n$

Z každého osudí vytáhneme r kuliček; očekávaný průměr počtu černých kuliček z i-tého osudí je tedy rp_i. Označme průměr pravděpodobností $p = \dfrac{p_1 + p_2 + \cdots + p_n}{n}$.

Vezmeme-li z každého osudí náhodný výběr r kuliček, bude celkový očekávaný průměr počtu černých kuliček mezi nr vytaženými $rp_1 + rp_2 + \ldots + rp_n = nrp$.

Jestliže je nrp očekávaný průměr počtu černých kuliček v nr tazích, je rp očekávaný průměr v r tazích, jež učiníme vždy z jednoho osudí náhodně vybraného. Tato hodnota

očekávaného průměru je totožná s očekávaným průměrem počtu černých kuliček ve výběrech r kuliček řady Bernoulliovy s konstantní pravděpodobností p.

Uvažujme nyní, jak velký bude rozptyl. Rozptyl ve výběru r kuliček z osudí O_i, kde pravděpodobnost černé je p_i, je dán výrazem rp_iq_i. Je to průměr čtverců odchylek od průměru výběrového z osudí O_i, jenž je rp_i. Hledáme však průměrnou čtvercovou odchylku od hodnoty rp místo od výběrového průměru rp_i. Chceme tedy stanovit obecný druhý moment kolem počátku v rp, který se podle rovnice (5) rovná druhému momentu kolem aritmetického průměru zvětšenému o čtverec rozdílu mezi průměrem a zvoleným počátkem. Je tudíž dán výrazem $rp_iq_i + (rp_i - rp)^2$.

Kdybychom vzali z jednoho osudí O_i takových náhodných výběrů na př. N, byl by ovšem očekávaný průměr součtu čtverců odchylek od rp větší N-krát, tedy

$$Nrp_iq_i + Nr^2(p_i - p)^2. \tag{83}$$

Utvoříme součet výrazů (83) pro všechna osudí, pak dostaneme

$$Nr\sum_{i=1}^{n} p_iq_i + Nr^2\sum_{i=1}^{n}(p_i - p)^2, \tag{84}$$

což je očekávaný průměr součtu čtverců odchylek od rp pro n osudí, z každého z nichž jsme vzali N náhodných výběrů rozsahu r kuliček.

Celkem máme Nn výběrů a poněvadž hledáme průměrnou čtvercovou odchylku od hodnoty rp, připadající na jeden výběr, kterou označíme S_L^2, musíme dělit jejich počtem poslední výraz (84), čímž dostaneme

$$S_L^2 = \frac{r}{n}\sum_{i=1}^{n} p_iq_i + \frac{r^2}{n}\sum_{i=1}^{n}(p_i - p)^2.$$

Součet v prvním členu na pravé straně rovnice však můžeme upraviti položíme-li $p_i = p + (p_i - p)$, a vzhledem

k $p_i + q_i = 1$ tedy $q_i = q - (p_i - p)$. Potom součin

$$p_i q_i = pq - (p_i - p)(p - q) - (p_i - p)^2$$

a součet jejich

$$\sum_{i=1}^{n} p_i q_i = npq - \sum_{i=1}^{n} (p_i - p)^2, \tag{85}$$

ježto

$$(p - q) \sum_{i=1}^{n} (p_i - p) = 0,$$

vzhledem k tomu, že

$$\sum_{i=1}^{n} (p_i - p) = 0,$$

neboť

$$p_1 + p_2 + \dots + p_n = np.$$

Na základě (85) bude tedy

$$S_L^2 = rpq + \frac{r^2 - r}{n} \sum_{i=1}^{n} (p_i - p)^2. \tag{86}$$

Označíme-li S_B^2 rozptyl výběru o rozsahu r z hypotetického souboru spočívajícího na schematu Bernoulliově s konstantní pravdě podobností p, která se rovná průměru daných pravděpodobností $p_1 + p_2 + \dots + p_n$, můžeme poslední rovnici psáti

$$S_L^2 = S_B^2 + \frac{r^2 + r}{n} \sum_{i=1}^{n} (p_i - p)^2$$

a vidíme, že rozptyl Lexisovy řady je větší než řady Bernoulliovy, spočívající na pravděpodobnosti p.

Příslušný výraz pro Lexisovy řady relativních četností dostaneme dělením pravé strany rovnice (86) čtvercem rozsahu výběru r^2, takže potom

$$\sigma_L^2 = \sigma_B^2 + \frac{1 - \frac{1}{r}}{n} \sum_{i=1}^{n} (p_i - p)^2;$$

pro velká r pak se užívá přibližně

$$\sigma_L'^2 = \sigma_B'^2 + \frac{1}{n}\sum_{i=1}^{n}(p_i - p)^2 \qquad (87)$$

a píšeme-li

$$\frac{1}{n}\sum_{i=1}^{n}(p_i - p)^2 = \sigma_p^2,$$

bude

$$\sigma_L^2 = \sigma_B^2 + \sigma_p^2. \qquad (88)$$

Druhý člen na pravé straně této rovnice se často nazývá **podstatnou komponentou kolísání**. Jiný způsob výkladu podává analysa rozptylu, o níž pojednáme později.

Při pozorovaných statistických řadách bude tedy směrodatná odchylka větší než směrodatná odchylka vypočtená z průměrné relativní četnosti znaku, která vystihuje náhodné kolísání bez vlivů rušivých.

Schema osudí s různým složením nám tedy zobrazilo rozptyl řad, který se tím více liší od normálního, čím se základní pravděpodobnosti těchto osudí od sebe více liší.

Rozptyl σ_L^2 však nemůžeme podle rovnice (87) počítati, ježto neznáme pravé hodnoty p_i resp. p, které se v ní vyskytují a musíme užíti přibližné hodnoty rozptylu

$$\sigma_f^2 = \frac{\sum_{i=1}^{n}(f - \bar{f})^2}{n}.$$

Stanovme tedy její očekávanou hodnotu.

Průměru $p = \frac{1}{n}\sum_{i=1}^{n} p_i$ odpovídá empiricky stanovený průměr $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$. Zavedeme si k řešení našeho úkolu identitu

$$f_i - \bar{f} = (f_i - p_i) + (p_i - p) - (\bar{f} - p),$$

takže její čtverec bude

$$(f_i - \bar{f})^2 = (f_i - p_i)^2 + (p_i - p)^2 + (\bar{f} - p)^2 + \\ + 2(f_i - p_i)(p_i - p) - 2(f_i - p_i)(\bar{f} - p) - \\ - 2(p_i - p)(\bar{f} - p).$$

Abychom stanovili očekávanou hodnotu $\mathfrak{E}(f_i - \bar{f})^2$ uvědomíme si, že

$$\mathfrak{E}(f_i) = p_i, \ \mathfrak{E}(f_i - p_i) = 0, \ \mathfrak{E}(\bar{f}) = p, \ \mathfrak{E}(\bar{f} - p) = 0,$$

podle rovnice (78) je

$$\mathfrak{E}(f_i - p_i)^2 = \frac{p_i q_i}{r}.$$

Poněvadž $\bar{f} - p = \frac{1}{n}\sum_{i=1}^{n}(f_i - p_i)$, je dále

$$\mathfrak{E}(f_i - p_i)(\bar{f} - p) = \mathfrak{E}\left[\frac{1}{n}(f_i - p_i)\sum_{i=1}^{n}(f_i - p_i)\right] = \\ = \mathfrak{E}\frac{1}{n}(f_i - p_i)^2 = \frac{p_i q_i}{rn},$$

neboť očekávaná hodnota ostatních $n - 1$ členů, v nichž se indexy liší, se rovná nule.

$$\mathfrak{E}(\bar{f} - p)^2 = \frac{1}{n^2}\mathfrak{E}\left[\sum_{i=1}^{n}(f_i - p_i)\right]^2 = \\ = \frac{1}{n^2}\left[\mathfrak{E}\sum_{i=1}^{n}(f_i - p_i)^2 + 2\,\mathfrak{E}\sum_{i=1,j=1}^{n}(f_i - p_i)(f_j - p_j)\right],$$

kde $i \neq j$; členy, kde by bylo $i = j$, se v tomto druhém součtu nevyskytují.

Očekávaná hodnota jednotlivých členů v druhém součtu se rovná nule a tedy i celého součtu, takže

$$\mathfrak{E}(\bar{f} - p)^2 = \frac{1}{n^2}\mathfrak{E}\sum_{i=1}^{n}(f_i - p_i)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\frac{p_i q_i}{r}.$$

Můžeme tudíž psáti celkem

$$\mathfrak{E}(f_i - f)^2 = \frac{p_i q_i}{r} + (p_i - p)^2 + \frac{1}{n^2} \sum_{i=1}^{n} \frac{p_i q_i}{r} - 2 \frac{p_i q_i}{r},$$

$$\mathfrak{E} \sum_{i=1}^{n} (f_i - \bar{f})^2 = \frac{\sum_{i=1}^{n} p_i q_i}{r} + \sum_{i=1}^{n} (p_i - p)^2 +$$

$$+ \frac{n}{n^2} \frac{\sum_{i=1}^{n} p_i q_i}{r} - \frac{2 \sum_{i=1}^{n} p_i q_i}{rn} =$$

$$= \frac{n-1}{n} \frac{\sum_{i=1}^{n} p_i q_i}{r} + \sum_{i=1}^{n} (p_i - p)^2.$$

Tento výsledek ještě můžeme upraviti, píšeme-li $p_i = p + (p_i - p)$, takže potom $q_i = q - (p_i - p)$, a součet je jako na str. 127.

$$\sum_{i=1}^{n} p_i q_i = npq - (p - q) \sum_{i=1}^{n} (p_i - p) - \sum_{i=1}^{n} (p_i - p)^2,$$

ježto prostřední člen se rovná nule, bude

$$\sum_{i=1}^{n} p_i q_i = npq - n\sigma_p^2$$

a konečný výsledek tedy je

$$\mathfrak{E} \frac{1}{n} \sum_{i=1}^{n} (f - \bar{f})^2 = \frac{n-1}{n} \frac{pq}{r} + \frac{n(r-1)+1}{nr} \sigma_p^2,$$

a při dostatečně velkém r se užívá obyčejně výrazu

$$\mathfrak{E}(\sigma_f^2) = \frac{n-1}{n} \frac{pq}{r} + \sigma_p^2. \qquad (89)$$

Jsou-li všechny pravděpodobnosti v základním souboru sobě rovny $p_1 = p_2 = \ldots = p_n = p$, pak je $\sigma_p^2 = 0$ a dostáváme již známý výsledek

$$\mathfrak{E}(\sigma_f^2) = \frac{n-1}{n}\,\frac{pq}{r}.$$

Obě hodnoty směrodatných odchylek se srovnávají utvořením podílu. Označme σ_f směrodatnou odchylku řady relativních četností pozorovaných ve studovaném souboru. Za předpokladu konstantní pravděpodobnosti p je pro příslušné rozdělení Bernoulliovo t. zv. teoretická hodnota směrodatné odchylky $\sigma_B = \left(\frac{pq}{r}\right)^{\frac{1}{2}}$.

Podíl $L = \frac{\sigma_f}{\sigma_B}$ se nazývá Lexisův poměr nebo koeficient. Také se nazývá v teoreticko-statistické literatuře koeficient divergence (podle Dormoye). Místo směrodatných odchylek rozdělení relativních četností bychom mohli použíti směrodatných odchylek rozdělení absolutních četností, neboť

$$\sigma_x = r\sigma_f \text{ a také } S_B = r\sigma_B.$$

Lexisův poměr je tím větší, čím se více odchyluje (diverguje) statisticky zkoumaný jev od dění náhodného.

Říká se, že řada pozorovaných relativních četností má rozptyl normální, je-li $L = 1$, nadnormální (supernormální), je-li $L > 1$ a podnormální (subnormální), je-li $L < 1$.

Vzhledem k (89) musíme tedy při zkoumání rozptylu srovnávati empirickou hodnotu σ_f^2 s výrazem $\frac{n-1}{n}\,\frac{pq}{r}$, při čemž za $\frac{pq}{r}$ musíme vzíti přibližnou hodnotu $\frac{\bar{f}(1-\bar{f})}{r}$.

Lexisův koeficient pak bude

$$Q^2 = 1 + \frac{n(r-1)+1}{nr}\sigma_p^2 : \frac{n-1}{n}\frac{pq}{r}, \qquad (90)$$

nebo přibližně

$$Q^2 = 1 + \sigma_p^2 : \frac{n-1}{n}\frac{pq}{r}. \qquad (91)$$

Máme tedy v koeficientu divergence důležitý prostředek k řešení nejvýznamnější úlohy statistiky, spočívající v zjištění, zda můžeme souditi na přítomnost změn v základních podmínkách výskytu znaku nebo na stále stejné, tedy konstantní, působení a složení základních podmínek.

Není sice absolutním kriteriem, ale dobrým vodítkem k posouzení kolísání výskytu znaku, jak u hromadných jevů fysikálních, tak sociálních.

V praktické statistice se velmi často vyskytují řady, které daleko přesahují míru očekávaného rozptylu. Za příklad si zvolíme statistiku úmrtí s nadnormálním rozptylem, jejíž rozbor provedeme podle Misesa k objasnění uvedené teorie.

Ve státě se 45 miliony obyvatelů byla na př. cifra úmrtnosti obyvatelstva, t. j. počet úmrtí připadající na 1000 obyvatelů v desítiletém období, v němž stejnoměrnost životní úrovně nebyla rušena nějakými pozoruhodnými vnějšími jevy, zaznamenána v těchto promilech 28,0, 27,8, 27,2, 27,5, 26,9, 27,2, 27,3, 27,4, 27,2, 27,6.

Tyto relativní četnosti naplňují údivem svou stálostí toho, kdo na ně pohlíží bez znalostí matematické teorie statistiky. Skutečně dřívější statistikové byli v úžasu nad mimořádnou stabilitou lidských poměrů, jevící se ve statistice. Dojdeme však ke zcela jinému závěru, vypočítáme-li skutečný rozptyl a srovnáme jej s očekávaným podle Lexisovy teorie.

Průměr uvedených deseti čísel je 27,41 promile, tudíž $\bar{f} = 0{,}02741$. Rozptyl pak dostaneme $\sigma_L^2 = 0{,}000\,000\,0949$. Očekávaná hodnota rozptylu řady Bernoulliho bude $\sigma^2 =$

$= \frac{\bar{f}(1-\bar{f})}{r} \cdot \frac{n-1}{n}$, kde koeficient $\frac{n-1}{n}$ vyplývá z teorie náhodného výběru podle (82), a hodnota p je nahrazena přibližnou hodnotou z pozorování, takže pro $r = 45\,000\,000$ (průměrný počet obyvatelstva v uvažovaném desetiletí) $\bar{f} = 0{,}02741$, $n = 10$ dostaneme $\sigma^2 = 0{,}000\,000\,000\,533$ a Lexisův poměr je $L = 13{,}34$. Přesahuje tedy skutečně pozorovaná směrodatná odchylka očekávanou teoretickou víc než 13kráte.

Naznačme, jak možno provésti rozbor tohoto výsledku. Lexisova teorie tu srovnává průběh roční úmrtnosti s deseti výběry, z nichž každý vznikl provedením 45 milionů tahů z osudí, v němž je stále mezi 100 000 kuličkami 2741 černých a 97 259 bílých. Kdyby na začátku každého z uvažovaných roků přišel každý obyvatel státu před toto osudí a vytáhl z něho svůj los života nebo smrti, museli bychom očekávati, že úmrtnost v tomto období vykáže rozptyl σ_B^2, který je 178krát menší než skutečně pozorovaný. Tento obraz nevystihuje hru o životě a smrti příléhavě, neboť ze zkušenosti víme, že mnohé příčiny smrti působí současně na řadu lidí, jako na př. nepříznivý vývoj povětrnosti v nějakém zimním nebo letním měsíci, endemické onemocnění atd. Vzhledem k tomu bychom se přiblížili skutečnosti lépe, kdybychom předpokládali, že za celý soubor přijde k osudí menší část a každý se otáže po osudu celé skupiny, kterou zastupuje.

Je zřejmo, že podle vzorce $\frac{pq}{r} \cdot \frac{n-1}{n}$ bude tato očekávaná hodnota rozptylu tolikrát větší, kolikrát bude počet nezávislých jednotlivých případů r menší. Kdybychom tedy předpokládali v našem případě, že pro každých 178 obyvatelů bude tažen společný los, který rozhodne o životě nebo smrti celé jejich skupiny, dostali bychom úplný souhlas mezi pozorováním a očekáváním. Zda lze v konkretním případě považovati vysvětlení silně nadnormálního rozptylu solidaritou jevů za případné, je třeba dále zkoumati. Bylo

by nutné, aby rozsah skupiny solidarity (178) zůstával zachován, když pozorujeme jiné analogické řady, na př. z jiných desítiletí. Kdyby to nebylo v dostačující míře splněno, bylo by třeba hledati jiné teoretické vysvětlení. V tomto případě lze je podati pomocí podstatné komponenty kolísání. Poněvadž se pravděpodobnost úmrtí rok od roku mění, jedná se o druhý typ Lexisovy řady, čili poměr černých a bílých kuliček je každý rok jiný. Potom, jak víme, je očekávaná hodnota rozptylu dána výrazem (89), čili k číslu nahoře vypočítaného rozptylu přistupuje další složka, která nezávisí na r, nýbrž jen na kolísání pravděpodobnosti od jednoho roku ke druhému. V tom, že podstatná komponenta kolísání nezávisí na r, v našem případě na počtu obyvatelstva, máme kontrolu teorie, neboť při kolísání pravděpodobnosti úmrtí, následkem hospodářských nebo klimatických poměrů v celém státě, musí se tato komponenta vyskytovati v přibližně stejné výši v jednotlivých větších oblastech státu.

3. Třetím typem jsou řady Poissonovy. Jejich schema si představíme tak, že náhodný výběr rozsahu r se skládá z prvků, z nichž každý byl vzat z osudí jiného složení; pravděpodobnost výskytu pozorovaného znaku je tedy u každého prvku výběru jiná, takže schema můžeme napsati takto

$$\begin{array}{c} O_1 \; O_2 \ldots O_r \\ \hline p_1 \; p_2 \cdots p_r \\ p_1 \; p_2 \cdots p_r \\ \cdots\cdots\cdots \\ p_1 \; p_2 \cdots p_r \end{array}$$

kde p je pravděpodobnost vytažení černé kuličky z osudí O_k.

Označíme-li průměr těchto pravděpodobností p, píšeme $p = \frac{p_1 + p_2 + \cdots + p_r}{r}$. Očekávaný průměr počtu černých kuliček ve výběru rozsahu r, jehož každý prvek je z jiného osudí, je rp a rovná se očekávanému průměru počtu černých

kuliček, bereme-li náhodný výběr rozsahu r z jednoho osudí o konstantní pravděpodobnosti p.

Odvodíme nyní rozptyl počtu černých kuliček v řadě Poissonově. Rozptyl pro osudí O_k je dán výrazem $S_k^2 = rp_kq_k$, kdyby byl celý výběr z něho vzat. Vezmeme-li jen jeden prvek z něho, položíme $r = 1$. Jsou-li pravděpodobnosti $p_1, p_2, \ldots, p_r$ na sobě nezávislé, pak platí věta o sčítání rozptylů (67), takže celkový rozptyl

$$S_P^2 = S_1^2 + S_2^2 + \ldots + S_r^2. \tag{92}$$

Dostáváme tudíž pro náš náhodný výběr

$$S_P^2 = p_1q_1 + p_2q_2 + \ldots + p_rq_r = \sum_{k=1}^{r} p_kq_k$$

a tento výraz opět upravíme tak, že položíme

$$p_k = p + (p_k - p)$$
$$q_k = q - (p_k - p),$$

takže

$$p_kq_k = pq - (p_k - p)(p - q) - (p_k - p)^2$$

a tudíž součet

$$\sum_{k=1}^{r} p_kq_k = rpq - \sum_{k=1}^{r} (p_k - p)^2,$$

neboť

$$\sum_{k=1}^{r} (p_k - p) = 0.$$

Pro rozptyl teoretického rozdělení počtu černých kuliček ve výběrech rozsahu r podle schematu Poissonova dostáváme tedy

$$S_P^2 = rpq - \sum_{k=1}^{r} (p_k - p)^2.$$

Je tedy rozptyl řady Poissonovy menší než rozptyl příslušné řady Bernoulliovy s konstantní pravděpodobností

rovnou průměru proměnné pravděpodobnosti:

$$S_P^2 = S_B^2 - \sum_{k=1}^{r} (p_k - p)^2. \tag{93}$$

Obdobnou rovnici pro rozdělení relativních četností lze snadno napsati jako v případě řad Lexisových

$$\sigma_P^2 = \sigma_B^2 - \frac{1}{r^2} \sum_{k=1}^{r} (p_k - p)^2. \tag{94}$$

(9,2) Koeficient nestálosti. Vedle Lexisova koeficientu zavedl Charlier koeficient nestálosti nebo disturbační, který rovněž měří vnější vlivy působící na změnu pravděpodobnosti v základním souboru. Definuje jej

$$\varrho = \frac{\sqrt{\sigma_L^2 - \sigma_B^2}}{p}. \tag{95}$$

Jeho přibližnou hodnotu dostáváme, klademe-li

místo σ_L^2 přibližnou hodnotu $\frac{\sum (f_i - \bar{f})^2}{n-1}$,

místo σ_B^2 „ „ $\frac{\bar{f}(1-\bar{f})}{r}$

a místo p „ „ $\bar{f}$.

Jako příklad si zvolíme poměr pohlaví živě-narozených dětí, což je velmi probádaným předmětem statistického šetření. Bylo mínění, že řady těchto čísel odpovídají poměrům náhodné hry o konstantní pravděpodobnosti, tedy řadě Bernoulliově. Přesto máme výsledky konkretních šetření, kde se objevuje rozptyl podnormální, což je v praksi statistické řídkým případem. Tak byly na př. ve Vídni (Wien) pozorovány ve 24 měsících let 1908 a 1909 tyto relativní četnosti chlapců mezi celkovým počtem živě narozených:

0,5223	0,5125	0,5141	0,5246	0,5126	0,5136
0,5187	0,5213	0,5105	0,5203	0,5124	0,5141
0,5143	0,5093	0,4904	0,5097	0,5140	0,5089
0,5129	0,5275	0,5178	0,5130	0,5177	0,5027

Jejich průměr je $\bar{f} = 0{,}514$ a rozptyl $\sigma_f^2 = 0{,}0000533$. Celkem se tam narodilo v té době 93 661 dětí, takže průměrně připadá na jeden měsíc $r = 3903$ dětí. Ve smyslu Lexisovy teorie se nyní ptáme, jaký je očekávaný rozptyl při konstantní pravděpodobnosti p, který by odpovídal 24 výběrům rozsahu 3903 z osudí stálého složení, kde mezi každým tisícem losů je 514 označeno znakem c (chlapec). Vypočítáme tedy tento rozptyl podle formule $\frac{pq}{r} \cdot \frac{n-1}{n}$, kde $n = 24$, $r = 3903$, $p = \bar{f} = 0{,}514$ a dostaneme $\sigma^2 = 0{,}0000613$, takže pro Lexisův poměr dostáváme $L = 0{,}93$, tedy rozptyl je podnormální. To nasvědčuje tomu, že případ neodpovídá osudí téhož složení, nýbrž jsou tu části obyvatelstva, jimž přísluší rozmanité pravděpodobnosti porodu se znakem c.

V podobném statistickém šetření, provedeném na př. ve Švédsku, byl zjištěn rozptyl poněkud nadnormální, kdežto na př. pro počet dvojčat v poměru k jednotlivým porodům se projevil rozptyl silně podnormální. Když byla pomocí čísla L konstatována existence rušivých vlivů na statistické řady, je pak úkolem statistikovým, pátrati po příčině poruch. Obecnou metodu k tomu dává teorie korelace. Na základě uvažování předložené řady lze dospěti jen k určitým závěrům o povaze rušivých vlivů, jimž je vystaven statisticky pozorovaný jev. Podle teorie Lexisovy vznikají tyto poruchy tím, že pravděpodobnost pro výskyt znaku se mění.

Úloha. Máme schema deseti osudí, z nichž každé obsahuje 15 kuliček, ale má postupně mezi nimi 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 bílých. Průměr pravděpodobností táhnouti bílou kuličku

je 0,5. Tvořme pokusem náhodné výběry po 10 prvcích tak, že z každého osudí vytáhneme jednu kuličku, pak je zase vrátíme do osudí a vytáhneme nových deset kuliček. Porovnejme potom čísla L pro 200, 500 po případě 1000 výběrů, abychom si pomocí tohoto kriteria zjistili stupeň shody, po případě neshody pozorování s tím, co očekáváme podle teorie.

LITERATURA.

1. *Janko:* Homogenita statistického souboru. (Statistický obzor, ročník XXI, čís. 9—10.)
2. *Krejčí:* Základy statistiky (1923).
3. *Horáček:* Základy statistiky (1935).

Učebnice vydané Ústředním statistickým úřadem:

4. Úvod do teorie statistiky (1926).
5. Základy teorie statistické metody (1929).
6. *Janko:* Základy statistické indukce (1937).

7. *Čuřík:* Počet vyrovnávací (1936).
8. Tabulky k numerickým methodám početním a matematické statistice. — Vydal spolek posluchačů pojistné techniky.
9. Českomoravské normy. (ČMN 2240 — 1940): Statistická kontrola jakosti (1940).
10. *Bydžovský-Teplý-Vyčichlo:* Aritmetika pro V.—VII. třídu středních škol. 6. vydání.
11. *Muk:* Aritmetika pro vyšší třídy gymnasií, reál. gymnasií a ref. reál. gymnasií.

OBSAH.

CESTA K VĚDĚNÍ SV. 22

Prof. Dr. J. Janko

Jak vytváří statistika obrazy světa a života - I. díl

Vyšlo roku 1942 nákladem Jednoty českých matematiků a fysiků v Praze

Tiskem knihtiskárny Prometheus v Praze

I. vydání - Cena brož. výtisku K 29,—


ZEMSKÁ UNIVERSITNÍ KNIHOVNA