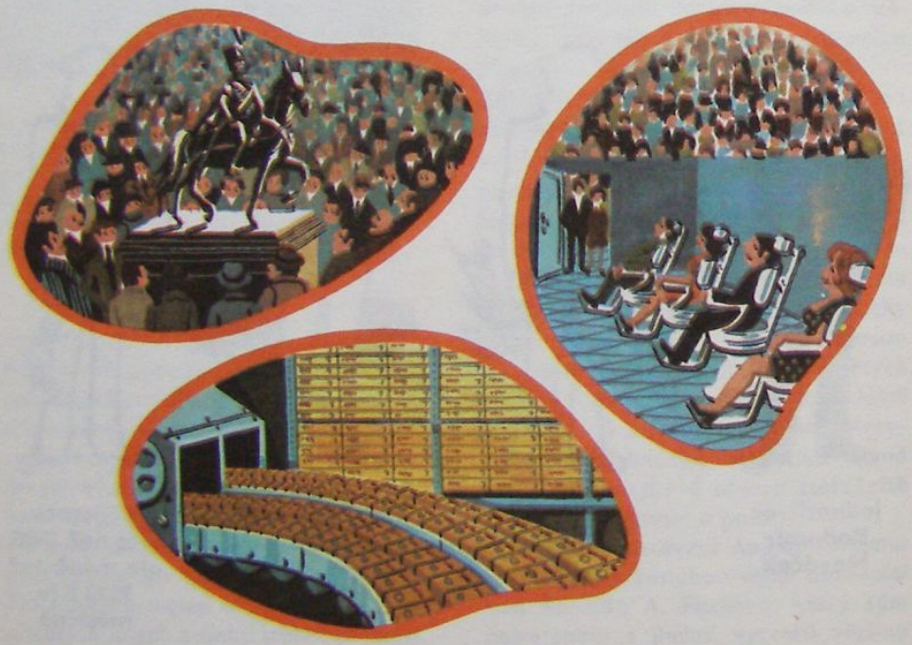
1.3 Základní soubor — rozdělení

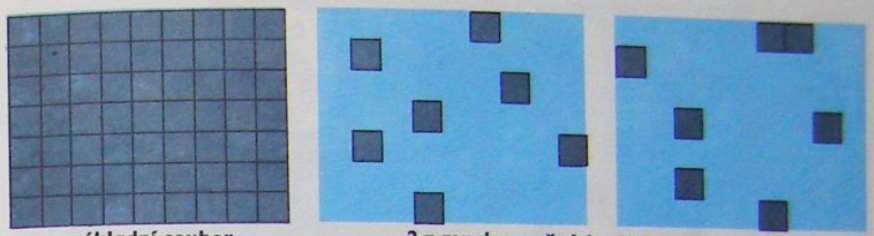
Hlavním obsahem této knihy je uvést a vysvětlit nejdůležitější statistické pojmy a metody. Nemělo by proto příliš mnoho smyslu, kdybychom je hned na začátku uvedli vyčerpávajícími definicemi, neboť se mohou stát srozumitelnými teprve ve vzájemných souvislostech. Základní myšlenku, která se jako červená nit vine dalším výkladem, chceme však nastínit hned na samém počátku.

Statistika se zabývá i ve své moderní podobě výběrové analýzy zásadně *hromadnými jevy*. Tím se nevylučuje, že ve výběru nejsou přesně zkoumány *jednotlivé věci nebo osoby*, avšak nikoliv proto, aby se zjistila jejich individuálnost, nýbrž proto, aby se zjistila *existence nebo neexistence nějakého znaku*, o němž se domníváme, že se vyskytuje i jinde, znaku, který je rozložen v základním souboru.

S následujícími dvěma slovy se budeme setkávat často. Jedno zní „*základní soubor*", druhé „*rozdělení*". Podívejme se nejdříve krátce na *základní soubor*. Je to většinou myšlenková konstrukce, která se nevyskytuje v přirozeném stavu, ale je formulována teprve v průběhu statistické činnosti. Takovým základním souborem může být „obyvatelstvo světa" nebo „obyvatelstvo



„Statistická masa", „základní soubor" může být tvořen podle libosti — musí však zahrnovat všechny jednotlivé prvky souboru: např. všechny ženaté muže v Brně, pacienty městské zubní kliniky v minulém roce nebo výrobu cigaret v závodě Y v měsíci dubnu.



základní soubor („soubor vyššího řádu")

2 z mnoha možných výběrových souborů („soubor nižšího řádu")

Je-li nemožné nebo příliš časově náročné, příliš drahé nebo neúčelné zachytit základní soubor ve všech jeho jednotlivých složkách, používá se výběrového souboru a libovolně se vyjme několik jednotek. Výběrový soubor poskytne pak více nebo méně spolehlivé závěry o základním souboru, a to podle rozsahu základního souboru, rozsahu výběrového souboru a rozdělení zkoumaných znaků.

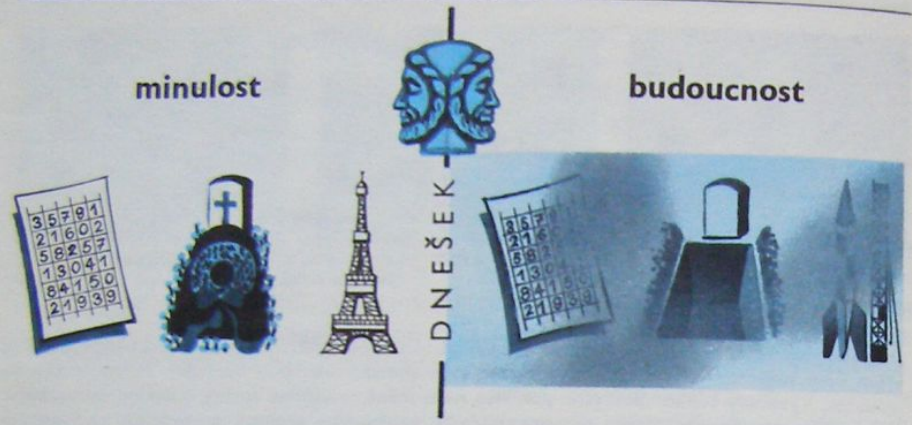
ČSSR k 1. 1. 1972" nebo také „ženatí muži v Plzni" anebo „samostatní zemědělci s ročním příjmem menším než 20 000 DM" či „pacienti městské zubní kliniky v Bordeaux" nebo „vozidla z roku 1970" anebo „cigarety vyrobené v měsíci dubnu závodem X a my Y".

Vymezení základního souboru není vždy docela bez problémů. Často je lákavé pokušení rozšířit jej více, než je zdůvodněno statisticky šetřeným materiálem. Když zkoumáme např. výběr 300 pacientů městské zubní kliniky v Bordeaux podle příjmu, rodinného stavu, počtu zubů, které ještě mají, nebo také podle barvy vlasů anebo tělesné váhy, dostaneme v případě, že výběr vzorku byl proveden statisticky správně, pravděpodobně dobrou představu řekněme o 10 000 pacientech, kteří navštívili tuto kliniku v uplynulém roce. Lze však použít tyto výsledky také pro pacienty zubní kliniky v Marseille nebo Lyonu, či dokonce v Edinburghu nebo Stuttgartu? To bude pravděpodobně nemožné bez dalšího šetření, zejména jde-li o zahraniční kliniku.

Teprve když jsme pomocí odpovídajícího výběrového šetření poznali do jisté míry např. základní soubor „pacienti zubní kliniky Stuttgart", budeme moci napříště — stále však ještě s náležitou obezřetností — aplikovat některé závěry šetření v Bordeaux také na Stuttgart.

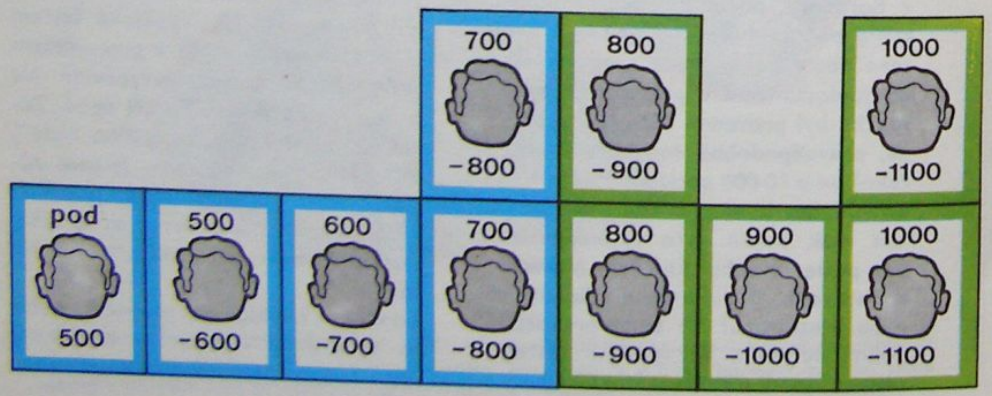
Zásadně však platí, že *výběrový soubor (vzorek) vypovídá jen o tom základním souboru, z něhož byl odvozen:* dotazníková akce v Mnichově poskytuje informace o názorech a poměrech v Mnichově, nikoliv však v Hamburku a už vůbec ne v Paříži. Výběrové šetření o mzdách vyplacených v papírenském průmyslu Hessenska nevypovídá nic o mzdách horníků v Porúří apod. *Základní soubor*, „soubor vyššího řádu", jak jej nazývá Anderson, *je onen základní soubor, v němž každý jednotlivý díl má stejnou naději dostat se do výběrového souboru* — „souboru nižšího řádu".

Co však je v základních souborech nebo ve výběrových souborech *rozděleno? Znak*, který je předmětem šetření nebo,

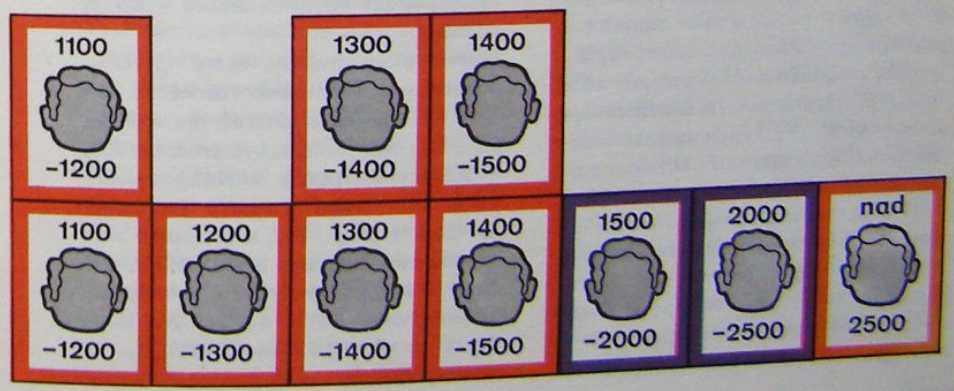


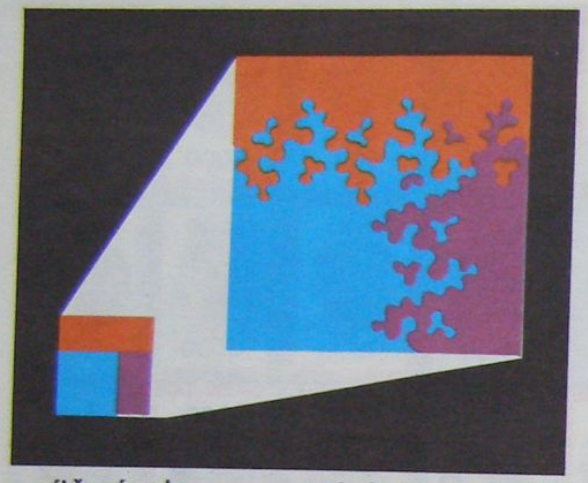
Statistické údaje jsou vždy „události", a proto skutečnosti z minulosti. Jejich projekce do budoucna je vždy zatížena nejistotou i při použití nejlepších postupů. Včerejší a dnešní čísla dovolují jen dohady a odhady čísel zítřejších, neumožňují však jejich přesný výpočet. Ze současné úmrtnosti se může pouze přibližně usuzovat o budoucí, z technických konstrukcí minulosti jen přibližně o konstrukcích v budoucnosti.

Téměř všechna statistická šetření se zásadně zakládají na pozorování nebo měření znaků v individuálních vyjádřeních: podle toho se zjišťují četnosti jednotlivých znaků. 25 osob dotázaných na svůj příjem se stane bezprostředně potom úplně neosobními „nositeli znaku" — nositeli, kteří se rozlišují jen více nebo méně četnými vyjádřeními znaku: dva z 25 vydělávají mezi 1500 a 2500 Kčs, pět z 25 vydělává méně než 800 Kčs — kdo jsou, jak se jmenují, zda mají příbuzné nebo dluhy, jsou nebo nejsou spokojeni, to je pro statistické zkoumání znaku „výše příjmu" prozatím bezvýznamné, což samozřejmě nevylučuje, aby v průběhu dalšího zkoumání souvislostí mezi příjmem a věkem, příjmem a školním vzděláním, příjmem a velikostí rodiny atd. se uvedené znaky zjišťovaly.



Znak: příjem





výběrový soubor neznámý základní soubor

Přesně známé je složení výběrového souboru. „Promítne-li" se však tento dílčí výsledek do neznámého základního souboru, přesnost se nutně ztrácí — zůstává jen odhad, ovšem v přesně propočitatelných mezích pravděpodobnosti a spolehlivosti.

přesněji řečeno, *četnost*, s níž se vyskytují určité charakteristiky znaku. Tak je třeba „rozdělen" důchod mezi příjemce důchodu. Někteří dostávají méně než 500 DM měsíčně, jiní mezi 500 a 550, další mezi 551 a 600 atd. až nad 1500 DM, kdy počet příjemců vyšších důchodů výrazně klesá (str. 26—27). „*Rozděleny*" jsou také četnosti, např. sled barev v ruletě. Velmi charakteristické rozdělení vykazují, jak ještě uvidíme, četná biologická měření, jako třeba tělesná výška nebo objem hrudníku: velmi vysokých a velmi malých lidí je málo, naproti tomu „průměrných" je velký počet. To je takzvané „*normální rozdělení*", které má ve statistice významnou roli. Vyskytuje se také „*rozdělení řídkých událostí*. Nikdo např. nemůže předvídat, kdo v příštím slosování bude mít 6 správných čísel, avšak docela dobře se dá předpovědět, kolik výherců přibližně bude.

Základní myšlenka, která je dnes základem značné části statistické činnosti, je tedy tato: má se propočítat (správněji: početním výkonem odhadnout), *jaké rozdělení určitého znaku je možno očekávat v základním souboru*, jestliže ve výběrovém vzorku jsem našel určité rozdělení.

Od průzkumu veřejného mínění ke kontrole výroby a od hospodářské prognózy k laboratornímu výzkumu se prostírá onen bezmála nekonečný prostor, v němž se na základě plánovaných experimentů, vzorků a dílčích poznatků usuzuje z *rozdělení* jednoho znaku *vzorku* na rozdělení stejného znaku v celém souboru.

Matematické metody, které vytvářejí početní souvislosti mezi vzorkem a celkem, jsou poněkud obtížné, ale nejdůležitější z nich mohou být srozumitelně vyjádřeny alespoň v hlavních rysech. Počtářské umění moderní statistiky, jak čtenář brzy uvidí, tvoří úplně svérázná matematika, spíše rafinovaná hra s pravděpodobnostmi a odhady než přesnost na několik desetinných míst, která se většinou s představou matematiky spojuje.

1.4 Pravděpodobnost a přesnost

Základem celé moderní statistiky je *počet pravděpodobnosti*, a to nikoliv jen z prostého a jasného důvodu, že zájem o údaje z minulosti vzniká především na základě zájmu o budoucnost; dílčí výsledky výběrových šetření mají především poskytnout pomůcku pro budoucí jednání.

Jde tedy vždy o odvození *závěrů* z *dílčích poznatků*, závěrů, které nesmějí nikdy být nesprávně chápány jako nesporné předpovědi, nýbrž jako předpovědi, které jsou vždy obklopeny pojmy jako „pravděpodobnost", „obor spolehlivosti", „maximální věrohodnost", „pravděpodobnost jistoty", „očekávaná hodnota", „rozptyl", „hypotéza" a „odhadovaná hodnota". Jednoznačně poznatelná může být nanejvýš minulost, nikdy budoucnost. „Poznání" (znalost) daného vzorku (výběrového souboru) však opravňuje k výpovědi o základním souboru, z něhož byl výběr proveden jen v přesně určených mezích. K tomu, aby se ze vzorků a podobných dílčích výsledků mohly odvodit závěry, jejichž vypovídací schopnost lze popsat, je nezbytně nutné, aby výběr ze základního souboru byl proveden *nahodile*, a nikoliv na podkladě záměrné volby. Jen v tomto případě je náhoda ponechána jaksi sama sobě, náhoda, kterou sice počet pravděpodobnosti dovede vymezit a podchytit, která je však falšována a křivena, jakmile se ji snažíme obejít subjektivně provedenou volbou (obr. na str. 31).

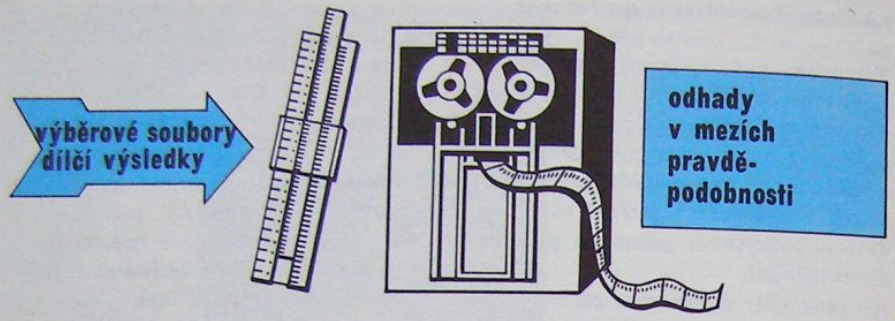
Z tohoto důvodu tvoří úvod do studia statistiky většinou studium počtu pravděpodobnosti a obráceně. Je velice obtížné vymezit přesnou hranici mezi počtem pravděpodobnosti a moderní statistikou. Neustále se objevují formulace, že statistika je koneckonců jen aplikovaný počet pravděpodobnosti. Je možné, že tato formulace není zcela správná, ale nikdo nepopře velmi úzkou souvislost mezi nimi.

Leonard Savage, jeden z nejvýznamnějších statistiků naší doby — a také, jak později ještě uvidíme, jedna z nejspornějších osobností — napsal k tomu ve svém díle „*The Foundations of Statistics*" (Základy statistiky): „Všeobecně se plně souhlasí s tím, že statistika se nějak zakládá na pravděpodobnosti. Co však se má rozumět slovem »pravděpodobnost« a jak je spojeno se statistikou, na to se názory tak různí a je tak malá možnost dorozumět se, jak sotva kdy od stavby babylónské věže."

V našem výkladu se ovšem nemůžeme podrobně zabývat těmito interními rozpory mezi veleknězi vědecké statistiky. Základní teoretická vysvětlení pravděpodobnosti jsme omezili na minimum, které je nezbytně nutné k porozumění statistické činnosti. Nicméně považujeme za vhodné poukázat již zde na toto úzké sepětí. *Propočet pravděpodobnosti, kvantitativní výpovědi o hypotézách, odhady a domněnky*, to jsou dnes hlavní úkoly statistiky. Matematicky dokonalé a přesné jsou jen *metody*, *výsledkem* jsou však odhady a pravděpodobnosti.

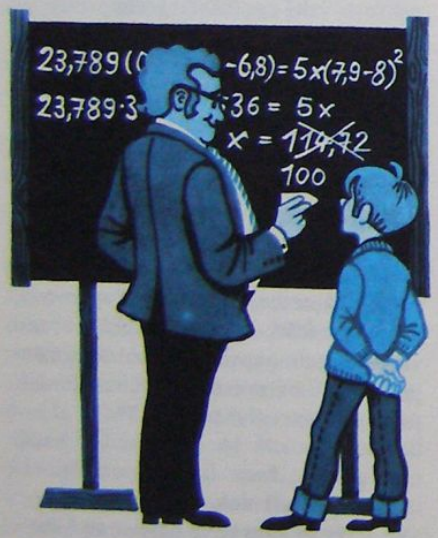
Zdánlivá přesnost statistiky klame jen laika. Odborník ví, že statistika je spíše *umění odhadu* a „*nauka o odhadu*" než *počtářská technika*. Ernst Wagemann, prezident Říšského statistického úřadu ve 20. letech, napsal: „Počtářské umění je jen služkou své mnohem vzdělanější paní, statistického odhadu."

Britský statistik M. Y. Moroney končí svou knihu „*Facts from Figures*" (Skutečnost z čísel) dokonce těmito slovy: „Ve škole se nám stále nalévá do hlavy,



Ani použitím všemožných pomůcek nelze dosáhnout toho, aby z výsledků výběrových souborů bylo možno učinit nesporné závěry o celkovém souboru. Je však zcela možné přesně zjistit kvalitu odhadu, např. tímto způsobem: „S 95% jistotou je tento odhad plus nebo minus dvě procenta správný."

že aritmetika je exaktní věda, že všechny úkoly v učebnicích mají svá správná řešení. A pokud slyšíme něco o odhadech, pak vždy s poznámkou, že přitom jde o hrubou přibližnost a že lze samozřejmě nalézt přesné řešení... To je ale žalostná příprava pro budoucí život. Snad až na pokladníka v bance, který počítá peníze jiných, je jinak úplná přesnost aritmetiky k ničemu... Učitelé by prokázali svým žákům velmi cennou službu, kdyby je učili kritickému přístupu k aritmetice. Měli by je učit umění škrtnout 114,72 a místo toho napsat 100..."



„Učitelé by se měli naučit umění škrtnout 114,72 a místo toho napsat 100," říká anglický statistik Moroney a vyjadřuje tím, že v denním životě se mnohem častěji setkáváme s odhady a přibližnými hodnotami než s úplně přesnými údaji.

Pro státovědnou statistiku 18. a 19. století měla pečlivá a přesná zjištění a úzkostlivě exaktní početní operace podstatný význam. Dnešní matematická statistika je naproti tomu převážně uměním určit, za *jak* nepřesné lze pokládat takové vypočtené údaje, jako je *pravděpodobný* výsledek, hypotéza či tvrzení.

Počátky tohoto poznání sahají daleko do minulosti. Quételet již před více než 100 lety varoval před přehnaným přeludem přesnosti. Laici, statističtí amatéři — a zvláště v minulých letech bohužel také statistikové z povolání — vždy znovu podléhali pokušení předstírat přesnost výsledků, která je popravdě prostě nedosažitelná. Oskar Anderson, jeden z prvních, kdo se v německé jazykové oblasti zasazovali o moderní statistiku, napsal ještě před necelými 40 lety na svou obranu toto: „Pro mne platí tyto dvě nezvratné zásady: 1. nemá praktický smysl chtít odvažovat fůru sena na chemicky přesných váhách; 2. není nic platné odhadnout vzdálenost mezi dvěma městy zhruba v tisíci krocích a pak k výsledku připočítat tloušťku městských hradeb v milimetrech." (Viz obr. na str. 32.)

Uvedená druhá „zásada" není ani dnes všude uznávána. Úcta před přesností aritmetiky, kterou už Moroney tak odsoudil, stále ještě silně působí. A pokud se skutečně někdy uvádí hrubý odhad, pak první, komu padne do ruky, s ním nakládá jako se svátostí. Tak můžeme číst, že 1609 km vysoko ve vesmíru nebo také 1609 km daleko od pobřeží se stalo to nebo ono: *odhad „1000 námořních mil"* byl přepočten s úzkostlivou přesností. Z téhož důvodu se katastrofy stávají často přesně 32 km od Tokia nebo San Franciska — *to zase, když se něco stane ve vzdálenosti asi 20 mil.*

Statistika opovrhuje tímto druhem počtářského umění — a právem. Není pochyb o tom, že statistická výpověď může být uvedena přesně až na desetinná místa, ale skutečně dobré statistice nikdy nebude chybět odkaz na to, jak velká je pravděpodobnost pro přibližnou správnost údajů. Jestliže statistik nechce říci: „asi 1600", řekne: „s 95% pravděpodobností ne méně než 1583 a ne více než 1631".

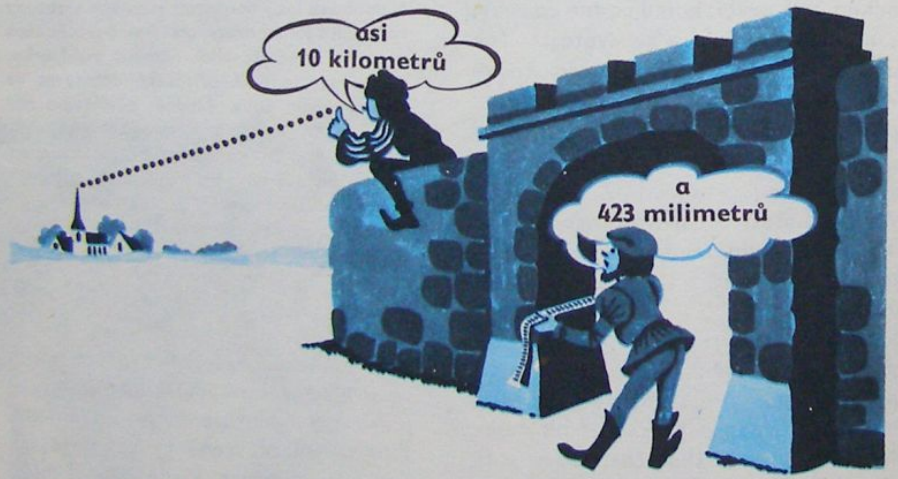
Zásadně musí každý jednotlivec mít stejnou šanci být pojat do výběrového souboru; jinak nelze použít postupů počtu pravděpodobnosti. Nemohu se tedy dotazovat jen lidí s klobouky nebo lidí s fotoaparáty, chci-li se dovědět něco o souboru návštěvníků náměstí sv. Marka. A nesmím se také přirozeně dotazovat ve stejnou denní dobu. Předně: návštěvníci náměstí sv. Marka vybraní pro výběrový soubor jsou jen návštěvníci náměstí sv. Marka, a nikoliv „Benátčané" nebo „turisté z Itálie" anebo „návštěvníci italských pamětihodností" či „hoteloví hosté v Benátkách".

Bylo by nesprávné jen *statistice připisovat k tíži takové zdůrazňování pravděpodobnosti a poukazování na nepřesnosti.* Zde se jen otevřeně vyslovuje to, co nechceme přiznat v denním životě, v běžném hovoru a často i ve vědě — totiž že i naše „ano" a „ne", dokonce naše „docela jistě" a „určitě ne" jsou zatížena často nejistotou a snad i zcela nepatrnou pochybností. Pro zjedno-

dušení a také proto, abychom se vyhnuli nenapravitelnému zmatku, nevšímáme si zpravidla náhod, překvapení a vzniklých nepravděpodobností. Matematická statistika to však činí viditelným, *vymezuje hranice pravděpodobnosti* a za to platí *zpochybněním přesnosti*.

Znovu však se musíme ptát spolu s Moroneyem: *k čemu je toto úsilí po přesnosti?* Pro správu železnic může být důležitá znalost počtu cestujících, kteří pojedou v příštím červenci z určitého města do Jugoslávie. K tomu ale plně postačuje znát počet cestujících v minulém roce zaokrouhleně na 1000 osob. Několik stovek více nebo méně už nehraje roli.

Ani vyčerpávající šetření není totiž bez vad. Nezřídka má větší vady než dobře a svědomitě provedené výběrové šetření. Jak problematickou se např. může stát při bližším pohledu i taková na první pohled prostá otázka, jako je počet cestujících v minulém létě z NSR do Itálie. Léto — co je to léto? Definujme: od začátku června do konce září. Co dál? Máme obejít všechny malé i velké cestovní kanceláře a sečíst záznamy cest do Itálie a prosit celníky na jižní hranici, aby se zeptali každého automobilisty na cíl jeho dovolené? Pak chybějí ještě jednotlivci cestující železnicí a ti, kteří jedou přes Francii. Ale cestující v osobních autech s paušální úhradou organizované cesty jsou již počítáni dvakrát. A co s těmi, kteří sice zaplatili let do Říma, ale odtud snad poletí dál do Káhiry nebo Madridu?



„Není nic platné, když vzdálenost mezi dvěma městy odhadneme a potom k výsledku připočteme tloušťku městských hradeb v milimetrech," řekl právem Oskar Anderson. Jestliže se k nějakému odhadu připočítá přesná míra, není tím celkový výsledek přesnější, nýbrž zůstává nadále odhadem.

Mají se mezi cestující do Itálie počítat také ti, kdo stráví dovolenou u Vrbského jezera v Korutanech a na tři dny si zajedou do Benátek?

Použije-li se údajů o přenocování a o přihláškách v cizině, budou čísla zase jiná, avšak nemusejí být lepší. Kolik majitelů soukromých pokojů raději své hosty nehlásí! Toho, kdo cestuje s obytným přívěsem, nelze skoro vůbec podchytit. Dotazníkovou akcí pomocí výběrového souboru několika tisíc občanů NSR se pravděpodobně dosáhne s mnohem menšími náklady správnějšího výsledku šetření než všemi uvedenými pokusy o vyčerpávající šetření.

Ještě většími zdroji chyb jsou údaje o tom, kolik DM nebo dolarů vydal v průměru turista ve Španělsku, Rakousku nebo Řecku. Všechny devizy totiž nenajdou cestu do Národní banky a dolary nemusejí vždy pocházet od Američanů a marky od Němců. Jestliže tedy čteme, že západoněmecký občan vydá v Itálii na osobu a den průměrně 3567 lir, je toto číslo především pouze odhadem a správně by mělo být vyjádřeno asi takto: „s 95% jistotou od 3500 do 3630 lir"; za druhé je toto číslo průměrem, který vzniká mechanickým rozdělením všech v Itálii vydaných DM na všechny německé cestující. Nicméně průměry mají ve statistice svůj velký — a často velmi špatně pochopený — význam. Věnujme jim proto pozornost hned na začátku.

Teorie pravděpodobnosti a statistika zacházejí s pojmy jako „jistota" a „určitost" pečlivěji než hovorová řeč. Pro budoucnost se nedají dělat žádné „absolutně jisté" předpovědi. Vědecký výhled do budoucnosti zná pojem „s nejvyšší pravděpodobností", nikoliv však „jistě".