Kapitola 5.

Kolik vran musíme pozorovat?

Když nevíš co děláš, zeptej se někoho, kdo to ví.

Jerry Pournelle, téměř v každém čísle magazínu BYTE

Tohle je opět kapitola o redukci informací. Je to kapitola přece jen radostnější než ty předchozí. Redukce populace na vzorek má dobře propracovanou teorii i dobře vypracované a spolehlivé recepty. Některé operace tu nejsou snadné, ale je mnoho lidí, kteří je znají a mohou nám poradit. Buďte tedy zadobře se statistiky.

Touto kapitolou vstupujeme do spíše technické oblasti výzkumu. K tomu nám může hodit dobrý pomocník. Dovolte, abych vám představil Dr. Watsona.

Dr. Watson je svým způsobem chytrý muž na systemizovaném místě pitomce. Je to někdo, koho každý profesor touží mít ve třídě. Doktor Watson vždycky navrhne nějakou, zdánlivě zřejmou, ale ve skutečnosti pitomoučkou odpověď, čímž umožní profesorovi nabídnout správnou odpověď, a tak se zaskvíti svojí moudrostí a učeností. Budeme služeb Dr. Watsona hodně používat.

5.1. Vzorek z nouze

Začneme spíše stupidní otázkou: "Kolik vran musíme pozorovat, abychom mohli říci, že všechny vrány jsou černé?" Odpověď je tak jednoduchá, že po ní nemusíme pátrat na konci kapitoly a přirozeně zní "Všechny!" Na druhé straně asi nikdo nikdy nepozoroval všechny vrány. Nezbývá nám nic jiného, než se spokojit s tvrzením, že "většina vran je černých". Opět je to něco, co už známe: redukovaná analýza reality vede k tvrzením pravděpodobnostního charakteru.

Skupiny, o které se v sociologickém výzkumu zajímáme, nejsou malé. V kvantitativní verzi výzkumu jsme schopni zkoumat celou skupinu jenom výjimečně. Pravidelně jediné sčítání lidu je studií celé populace. Většinou studujeme jen některé členy skupiny a doufáme, že naše závěry budou aplikovatelné i na ostatní, na ty nestudované. To nás přivádí k dvěma základním termínům, které potřebujeme pro tuhle kapitolu: populace a vzorek (výběrový soubor). Jejich definice je jednoduchá:

VZOREK:	skupina jednotek, které skutečně pozorujeme
POPULACE	(neboli základní soubor) je soubor jednotek, o kterém předpokládáme, že jsou pro něj naše závěry platné

Náš stěžejní úkol je najít postup, aby výsledky, které získáme na vzorku, byly co nejvíce podobné těm, které bychom získali na celé populaci. První věc, která nám přijde na mysl, je, snažit se mít vzorek co největší. Ale naše následující pravdivá pohádka nám ukáže, že to není jen tak:

Pohádka pro odrostlejší děti 8.

O hodně velkém vzorku, aneb Jak to nevyšlo

Byl jednou v Americe velice rozšířený týdeník, který se jmenoval Literary Digest. Byl u svých čtenářů hodně oblíben. Byl proslulý také tím, že spolehlivě předpovídal výsledky presidentských voleb. Jeho předpovědi byly založeny na obrovském vzorku dvou milionů voličů. (Dnes jsou podobné předpovědi založeny na vzorku tisíckrát menším.) Vzorek byl zkonstruován z mnoha zdrojů. Literary Digest si opatřil adresy voličů z celých USA. Používal pro to zdroje jako telefonní seznamy, městské adresáře, adresy držitelů řidičských průkazů, členské seznamy organizací, seznamy předplatitelů novin a časopisů atd.

Předpovědi byly přesné a úspěšné ve volbách 1920, 1924, 1928, 1932, a pak přišly volby v roce 1936. Literary Digest předpověděl, že presidentský kandidát Landon porazí Roosevelta rozdílem 14%. Přišel volební den a s ním i konec slávy Literary Digestu: Franklin Delano Roosevelt zvítězil drtivou většinou.

Cvičení 4.1.

Reprezentoval vzorek použitý Literary Digestem dobře celou populaci voličů v USA?

To nebylo tak těžké, že? Trochu složitější je otázka, jak je možné, že vzorek, který prakticky vyloučil z výzkumu voliče náležející k nižším sociálním třídám, fungoval dobře v předchozích volbách? Klíčem k řešení je rok: v roce 1935 vrcholila v USA hospodářská krize, a to vedlo k ostré polarizaci podle vertikální stratifikační osy. Předtím sociálně ekonomický status nehrál příliš důležitou roli v otázce volebních preferencí. Daleko větší úlohu hrály takové faktory jako náboženství, zeměpisná poloha atd. Krize to všechno změnila: sociální status začal hrát důležitou funkci. Pravděpodobně nejdůležitější bylo to, že krize přivedla k volebním urnám příslušníky nižších sociálně ekonomických vrstev, kteří předtím příliš často nehlasovali. Můžeme tedy říci, že v letech 1920-1932 předpovědi Literary Digestu vyšly jenom náhodou. Abychom byli schopni z chování vzorku předpovídat chování populace, **musí struktura vzorku imitovat složení populace tak přesně, jak je to jen možné.**

Dr Watson: Ale to je přeci docela lehké! Když je v populaci řekněme 51% žen, tak vyberu také 51% žen do vzorku, a když je v populaci 12% osob nad 65 let věku, vyberu také stejné procento starých osob do vzorku, atd.

Tentokrát má Dr. Watson pravdu. Technika konstrukce vzorku, tak jak ji popsal,se opravdu používá. Říká se tomu **kvótní výběr.**

> Kvótní výběr imituje ve struktuře vzorku známé vlastnosti populace.

Bohužel má tato technika některé nepříjemné vlastnosti. Jedna z nich souvisí se slovem "známé" v naší definici. Pro většinu populací není problém zjistit jejich skladbu podle

pohlaví, věku, vzdělání, povolání atd. Lze si snadno představit problém, pro který jsou důležitější jiné vlastnosti, takové, o kterých běžná statistická šetření údaje neshromažďují (kupř. věk, ve kterém se respondent poprvé zamiloval).

Na další problém snadno přijdete sami:

Cvičení 4.2.

Navrhněte prosím, kritéria pro konstrukci kvótního vzorku pro populaci veksláků.

Kvótní výběr může být použit jen na populaci, o které jsme dobře informováni, a to zdaleka není každá populace. Další obtíž je spojena s praktickou stránkou výběru přímo v terénu. Poslední krok obvykle závisí na tazateli, který vybírá jedince podle dané instrukce. Taková instrukce by mohla vypadat třeba takto:

<u>Jméno tazatele:</u> *Dr. Watson*

Respondent č.1.

muž, věk 30-40, dokončené středoškolské vzdělání, povoláním úředník, ženatý, ale bezdětný, bydlící v rodinném domku, žijící v našem městě alespoň 5 let, ale který se narodil v obci pod pětset obyvatel...

Respondent č.2.

žena, věk 60-65, alespoň s dokončeným základním vzděláním, důchodkyně, která pokud byla ještě ekonomicky aktivní, měla dělnické povolání, která žije sama, v bytě alespoň o dvou místnostech a bydlí od narození v našem městě...

Tak, to si od nás Dr. Watson opravdu nezaslouží. Umíte si představit, na kolik dveří by musel zaklepat, než by našel osoby, odpovídající zmíněným charakteristikám. Třeba by je nenašel vůbec, možná, že vůbec neexistují. Ve skutečnosti je instrukce v kvótním výběru mnohem skromnější. Navrhuje jen několik málo proměnných, takových jako pohlaví, věk a povolání. Lokalita a typ obce je obvykle dán působištěm tazatele. Jinak nejsou tyto proměnné vázány do určitých kombinací. Instrukce by mohla znít takto: "Hovořte s deseti osobami, z toho se šesti ženami a čtyřmi muži. Vyberte 3 osoby ve věku pod 20 let, 5 ve věku 21-50.." O ostatních, pro nás třeba daleko důležitějších proměnných můžeme jenom doufat, že budou ve vzorku dostatečně správně reprezentovány.

Dr. Watson: Co si s tím ale počneme?

Odpověď nám nabízí titul následujícího paragrafu.

5.2. Hodíme si korunou aneb Pravděpodobnost pro Dr. Watsona

Představme si, že máme velikou krabici, plnou kuliček, a že všechny kuličky jsou zelené. Dobře krabicí zatřepeme a poslepu vybereme jednu kuličku. Jakou máme šanci, že vybraná kulička bude zelená? To byla ale pitomá otázka, že ano? Tak si teď zkusme něco trochu složitějšího: Máme teď jinou populaci kuliček, sestávající ze zelených a červených kuliček. Těch zelených je 80% a těch červených je ovšem 20%. Ale počkejte, já se vás nebudu ptát, jaká je pravděpodobnost, že si náhodně vyberete červenou kuličku. To byla otázka jen o málo méně pitomá, než ta první, a všichni víme, že ta pravděpodobnost je 20%, a chceme-li to vyjádřit učeněji, můžeme říci, že $p = 0{,}20$.

My tu máme jiný úkol: zjistit, jaká je skladba populace, aniž bychom prohlíželi všechny kuličky. Jinými slovy, hledáme metodu, jak vytvořit vzorek, který by dobře reprezentoval celou populaci kuliček. Můžeme zkusit třeba toto: Opět začneme tím, že krabicí dobře zatřeseme. To není vtip, to je opravdu nutné: **každá kulička musí mít stejnou pravděpodobnost, že bude vybrána.** (Co kdyby všechny červené kuličky byly navrchu?) a teď vybereme poslepu 10 kuliček. Uvidíme třeba, že jsme vybrali 6 červených a 4 zelené. To je dost daleko od dobré reprezentativity. Perfektní vzorek by měl přeci obsahovat 20% červených a 80% zelených. Tedy vybereme opět poslepu dalších deset kuliček. Třeba 6 z nich bude zelených a 4 červené. Přidáme je k našemu původnímu vzorku. Nový, větší vzorek sestává z 10ti červených a 10ti zelených kuliček. Teď bychom odhadli, že v populaci je stejné procento červených, jako zelených kuliček. To ještě není vůbec dobré. Museli bychom tedy pokračovat, přidávat další a další kuličky. Brzy bychom zpozorovali zajímavou věc:

Sostoucí velikostí vzorku se rozdíl mezi strukturou populace a vzorku zmenšuje.

Nejdříve rychle, pak pomaleji a pomaleji. Úplné shody mezi strukturou populace dosáhneme teprve tehdy, když jsme zahrnuli všechny elementy populace do vzorku.

Dr.Watson: "Ale to je všechno nesmysl! Když je to pravda, jak je potom možné, že obrovský vzorek použitý Literary Digestem vedl k tak nesprávným výsledkům?"

Asi už víte, co bychom mohli odpověď na tuhle námitku: "Ale to je přece elementární, Watsone. Ti lidé z Literary Digestu zapomněli pořádně zatřást krabicí." Voliči z nižších socioekonomických vrstev měli mnohem menší šanci být vybráni do vzorku, než voliči ze středních a vyšších vrstev, což dramaticky zkreslilo výsledky.

My jsme tu totiž, aniž bychom o tom věděli, vytvořili náhodný vzorek "populace" kuliček. A náhodný vzorek, to je aristokrat mezi vzorky; má mnoho jedinečných, a pro nás důležitých, vlastností. Všechno, co budeme v tomto odstavci probírat, se týká jenom vzorků, které byly vytvořeny opravdu náhodným výběrem. Termín "náhodný" neznamená výběr nazdařbůh. I když náhodný výběr může být, jak brzy uvidíme, technicky velmi obtížný a často i nemožný, jeho definice je jednoduchá:

Náhodný (pravděpodobnostní) výběr je takový výběr, ve kterém každý element populace má stejnou pravděpodobnost, že bude vybrán do vzorku.

To se lépe řekne než se to udělá. Ale dovolte, abych vás ještě dříve než budeme mluvit o řadě trampot, dobře naladil popisem pozoruhodných vlastností náhodného vzorku. Snad nejdůležitější z nich, alespoň pro nás sociology - statistik by s námi možná nesouhlasil - je tato vlastnost:

Náhodný vzorek reprezentuje všechny známé i neznámé vlastnosti populace.

A ještě dříve, než Dr. Watson začne namítat, uveďme si jednoduchý příklad. Máme teď novou populaci kuliček. Jsou opět červené a zelené. Ale mají ještě jednu zajímavou vlastnost, o které my nevíme: Jsou duté a uvnitř každé je malý papírek a na každém tom lístku je něco napsáno. (Znáte "fortune cookies" z čínských restaurací?) Třeba nějaké neslušné slovo. Když jsme vybrali dobrý náhodný vzorek kuliček, budou reprezentovat celou populaci kuliček nejen vzhledem k distribuci barev, ale i vzhledem k distribuci neslušných slov, i když o tom nevíme a třeba nikdy nebudeme vědět. Uveďme si jiný, užitečnější příklad. V náhodném vzorku obyvatelstva hlavního města Prahy budeme mít slušnou reprezentaci populace vzhledem k věku, pohlaví, vzdělání, povolání, politické orientaci, vzhledem ke všem postojům, ale i reprezentaci třeba vzhledem k oblíbeným jídlům, počtu zubních kazů, věku, kdy se lidé poprvé zamilovali, množství vypitého piva, počtu milenek, počtu veksláků, peněžní hodnotě nakradeného zboží, číslům bot, prostě vzhledem ke všemu. To neznamená, že tohle všechno budeme schopni měřit, to je jiný problém. Ale znamená to, že ať už je naším cílem cokoli, víme, **že proměnné, které jsou pro nás relevantní, budou mít v našem vzorku podobnou distribuci, jaká existuje v celé populaci a naše závěry jsou tedy na tuto populaci aplikovatelné.**

Náhodný výběr má ještě jednu pozoruhodnou vlastnost:

U náhodného vzorku jsme schopni odhadnout, jak se vzorek liší od populace.

Jinými slovy, jsme schopni určit, jak dobrý je náš vzorek. Teď je na čase naučit se několik slov z odborné hantýrky, jednak abychom mohli oslnit přátele, jednak abychom rozuměli správně významu publikovaných statistických dat. Podívejme se na následující tabulku:

Tabulka 5.1.

Velikost vzorku a konfidenční interval

na 95% hladině významnosti pro alternativní znaky při distribuci 50:50

Velikost vzorku	Konfidenční interval
100	± 10%
400	± 5%
1600	± 2.5%

Adaptováno z Babbie: Social Research for Consumer, 1982

To vypadá dost učeně, že? Ale nebojte se. Pochopit princip, a vědět jak se taková věc aplikuje, není těžké. Trochu obtížnější je statistické zdůvodnění. Ale takové vysvětlení necháme pro někoho jiného, kdo vás uvede do zajímavého světa **skutečné** statistiky.
Řekněme, že jsme vybrali náhodně 400 kuliček a zjistili jsme, že ve vzorku (neboli ve výběrovém souboru) je 78% zelených kuliček. Protože jsme nevybrali všechny kuličky, musíme předpokládat, že jsme se dopustili určité chyby, že pozorovaná relativní četnost zelených kuliček ve vzorku se liší od procenta, které skutečně existuje v celé populaci (základním souboru). My však potřebujeme vědět, jak moc se mýlíme. A v tom nám pomůže ta nepřátelsky vyhlížející tabulka. Pozor! **Tahle tabulka je jen ilustrací a platí jen tehdy, je-li v populaci právě tolik zelených jako červených kuliček. Platí jen pro alternativní (binomické) proměnné,** to je pro takové znaky, které mají jen dvě kategorie, jako ANO a NE. V našem případě, zelená a "nezelená" kulička.

Velikost našeho vzorku je 400 a této velikosti vzorku odpovídá konfidenční interval (interval spolehlivosti) 5%. Odečteme tedy tuto hodnotu od pozorovaných 78% a dostaneme tedy 73%. Pak ji opět přičteme k pozorované hodnotě a dostaneme horní mez, a teď víme, že skutečná proporce zelených kuliček v celé populaci je mezi 73 a 83%.Jenomže to nevíme docela určitě, vždyť jsme nepozorovali všechny kuličky. Teď se dostáváme k tomu poněkud kryptickému výrazu v podtitulu naší tabulky: **hladina významnosti**.

V našem případě to znamená, že skutečná proporce, která existuje v populaci, se nalézá s 95% pravděpodobností uvnitř vypočítaného intervalu spolehlivosti. Kdybychom vytvořili 100 vzorků obdobné velikosti, jen v 5 vzorcích by bylo možné, že skutečná proporce zelených kuliček leží pod nebo nad vypočítaným konfidenčním intervalem. O tom, jakou hladinu zvolit, rozhodne výzkumník, a podle tohoto rozhodnutí je interval vypočítáván. Toto rozhodnutí je svobodné ovšem jen z hlediska statistické teorie; ve skutečnosti je vázán míněním, přijatým v příslušné vědecké komunitě. V sociologii je to obvykle 95 nebo 98%. (Vidíte, i v sociologii máme malý kousek paradigmatu.)

A teď se podívejme, jak by se takový interval **mohl** vypočítat. Není to tak, jak se to opravdu dělá. Ve skutečnosti neznáme distribuci proměnné, která existuje v populaci. Ale náš popis výpočtu nám dá alespoň nějaký vhled do logiky, která je skryta za pozoruhodnými vlastnostmi náhodného výběru. Protože jsem vám slíbil, že v naší knize nebudou (skoro) žádné vzorečky, popíšeme si výpočet slovně. Nejdříve musíme vypočítat veličinu, která má opravdu zajímavé vlastnosti a které se říká **směrodatná chyba**. Uvidíte, že je to nejen snadné vypočítat, ale také, že není těžké rozumět většině kroků v tomto výpočtu.

Výpočet směrodatné chyby:

CO UDĚLÁME	CO TO ZNAMENÁ
Nejdříve vynásobíme proporci zelených kuliček v populaci proporcí červených. Tato proporce musí být vyjádřena jako desetinný vzorek, ne v procentech. (Tedy, kdyby v populaci bylo 50% červených a 50% zelených budeme počítat 0.5 krát 0.5.)	Homogenita vzorku má vliv na velikost chyby. Čím nerovnoměrnější je distribuce ve vzorku, tím menší bude chyba a tím užší bude interval spolehlivosti. Kdyby na příklad v populaci bylo 90% zelených kuliček a velikost vzorku by byla 100, vypočítaný konfidenční interval by byl ± 6%. Kdyby ve stejně velkém vzorku byl stejný počet zelených jako červených kuliček, konfidenční interval by byl mnohem širší: ± 10%
Vypočítaný násobek vydělíme velikostí vzorku.	Čím větší vzorek, tím menší je směrodatná chyba a tím užší bude konfidenční interval. V případě, že by v populaci byla stejná proporce zelených a červených kuliček, ve vzorku 100 pozorování, by interval byl ± 10%; ve vzorku 400 pozorování by byl mnohem užší: ±5% a ve vzorku 1000: ±3%.
Nakonec vypočítáme druhou odmocninu z výsledku dělení.	To je transformace do čísla zajímavých vlastností. Ti, kdo jsou trošku seznámeni se statistikou, vidí už teď souvislost s konceptem směrodatné odchylky. My ostatní to pochopíme trochu lépe, až budeme mluvit o směrodatné odchylce v naší statistické kapitole.

A teď nám už zbývá jen jedno. Rozhodnout se, jakou hladinu významnosti chceme přijmout, a pak vypočítat interval spolehlivosti.

Směrodatná chyba má jednu pozoruhodnou vlastnost: do intervalu vymezeného ±1 standardní chybou od hodnoty pozorované ve vzorku připadne správná hodnota, existující v populaci, přibližně v 68 případech ze sta. Tak bychom dostali interval spolehlivosti na 68 % hladině významnosti. To ovšem není zdaleka dost vysoká pravděpodobnost. Abychom vypočítali interval spolehlivosti na úrovni, jaká je vyžadována v našem oboru, musíme přičíst a odečíst směrodatnou chybu dvakrát. Jinými slovy: **interval spolehlivosti na 95% hladině významnosti je dán rozmezím ± 2 směrodatné chyby od hodnoty, naměřené ve vzorku.** Rozmezí ±3 směrodatné chyby nám definuje ještě mnohem striktnější interval na hladině 99.9%. Ten je užíván zejména v přírodních vědách.

A teď už víme dost, abychom mohli představit další, opravdu překvapivou vlastnost náhodného výběru:

Velikost směrodatné chyby, a tedy i konfidenční interval (interval spolehlivosti) nezávisí vůbec na velikosti populace.

Jedině velikost vzorku a jeho homogenity ovlivňují velikost chyby.

Dr.Watson:

Počkejte, počkejte! Chcete mi namluvit, že řekněme vzorek 300 respondentů vykáže stejnou chybu, když reprezentuje populaci továrny s 800 dělníky, jako stejně velký vzorek, který reprezentuje město s 50.000 obyvatel, nebo dokonce zemi s 200,000.000 občanů? Já tomu prostě nevěřím!

Neuvěřitelné, a přece je to pravda, pokud ovšem distribuce zkoumané proměnné je ve všech těch populacích stejně homogenní. A pokud mi ještě nevěříte, podívejte se znovu na popis výpočtu směrodatné chyby. Najdete tam zmíněnou proporci zelených a červených kuliček, velikost vzorku a to je vše. Ani zmínka o populaci.

To, co víme, by nám mohlo dát dostatečnou informaci, abychom mohli **navrhnout velikost vzorku, jakou potřebujeme vzhledem k velikosti chyby, jakou jsme ochotni riskovat.** V praxi to však není snadné: pro výpočet směrodatné chyby potřebujeme znát homogenitu populace vzhledem k našim proměnným, rozptyl těchto proměnných. Většinou tuto znalost nemáme. Existují sice techniky, které nám umožní tuto informaci odhadnout, ale tyto techniky jsou buďto nákladné nebo nepřesné.

A tak v tvrdé praxi denního života výzkumníka spoléháme na zkušenost a na zdravý rozum. Můžeme se třeba zamyslit nad tím, které kombinace proměnných jsou pro nás nejdůležitější. Představíme si kolik polí bude mít tabulka (nebo tabulky) a navrhneme, kolik pozorování musí každé pole v těchto tabulkách obsahovat - prázdná pole, nebo pole s málo pozorováními mohou podstatně zkreslit výsledky statistické analýzy. Zaměřme se raději na dost vysoké minimum; někdy navrhovaný průměr 10 pozorování na jedno pole tabulek může být nezdravě optimistický. Data ve skutečnosti nebudou do všech polí rozdělena rovnoměrně; některá pole budou přeplněna a jiná téměř prázdná. Nadto v každém výzkumu máme mnoho proměnných, s různým počtem kategorií, někdy nevíme předem, které kombinace proměnných přinesou nějaké zajímavé výsledky, a tak si zaslouží hlubší analýzy atd. Zkrátka, teoretizování o velikosti vzorku patří spíše na stránky učebnic než do praxe sociologického výzkumu. Tam aplikujeme následující, velice nevědecké, ale velice praktické pravidlo: **Snažme se vytvořit**

co největší vzorek, jaký nám naše časové a finanční podmínky dovolují; ne však za cenu vážného narušení pravidel náhodného výběru. Doba pro aplikaci naší znalosti o intervalech spolehlivosti přichází v praxi teprve v etapě statistické analýzy sebraných dat. Pak je to ovšem velice důležité.

A teď ještě jedno důležité varování:

> Velikost směrodatné chyby se týká jen zkreslení, vyvolaného rozdíly mezi vzorkem a populací. Nevztahuje se, bohužel, na zkreslení vyvolané jinými typy redukce a transformace informací. Tato zkreslení jsou pro nás většinou mnohem nebezpečnější a my nemáme žádný nástroj, jak měřit velikost těchto omylů.

5.3. Jak správně házet korunou

Dr. Watson:

Já už vidím, že náhodný výběr je výborný. Hned to začnu používat. Vždycky jsem chtěl vědět, co si lidé v Praze myslí o mojí politické straně. Hned začnu pracovat na hypotézách a otázkách pro rozhovor. Od pondělí budu každé dopoledne na Václaváku a budu se vyptávat náhodně vybrané osoby...

Pokud náš pošetilý přítel doufá, že jeho výsledky budou reprezentovat mínění pražské populace, je ještě mnohem pošetilejší, než jsme si mysleli. Víme přece, že při náhodném výběru každý člen populace musí mít stejnou pravděpodobnost, že bude vybrán. Watsonův vzorek by byl silně zkreslený.

Cvičení 5.3.

Navrhněte prosím, jak by se Watsonův vzorek lišil od pražské populace.

Tedy jasně vidíme, že tento vzorek by snad mohl být reprezentativní pro populaci definovanou asi takto: osoby, které se nacházejí na Václaváku ve všední den dopoledne, v dané roční době. Pro nějaké speciální účely by mohla být taková populace zajímavá: kupř.

pro plánování obchodních strategií pro obchody na Václaváku, rozhodně však ne pro problémy spojené s politickou orientací obyvatel. Ale i tak by byla **náhodnost**, a tedy i reprezentativnost takového výběru problematická. Dr. Watson, protože je v podstatě konzervativní, by se mohl ostýchat oslovit méně konvenčně oblečené osoby. Kdyby takový výběr prováděl můj syn, půvabné mladé ženy by byly ve vzorku přereprezentovány. Kdybych prováděl výběr já, pak by byly podreprezentovány, protože jsem stydlivý. Ono se vůbec zdá, že lidská mysl není schopna pracovat opravdu náhodně.

Můžeme si to dost snadno vyzkoušet. Požádejte větší skupinu lidí - třeba třídu studentů - aby každý napsal na kousek papíru jakékoliv číslo mezi 1 a 10. Bez dlouhého přemýšlení musí napsat to, co jim přijde na mysl. Je-li skupina dost velká, je vysoká pravděpodobnost, že číslo 7 bude mít daleko nejvyšší frekvenci. Proč, to nevím, a předem můžete zavrhnout teorii vlivu sedmy v naší mariášnické kultuře; v Kanadě to funguje také, a jak! Snad to má něco dělat s tradiční mystikou čísel, ale v každém případě to krásně dokumentuje, že náš mozek je velice špatným generátorem náhodnosti. Musíme jej nahradit něčím neosobním. Hodit si korunou? Zatřepat krabicí?

Pomůcky, které v praxi při výběru náhodného vzorku používáme, skutečně imitují takové mechanismy. Mohli bychom třeba napsat jména všech členů populace na papírky, dát do klobouku, kloboukem pořádně zatřepat a pak poslepu vytáhnout tolik papírků, kolik osob potřebujeme do vzorku. Ovšem většinou by to musel být pěkně velký klobouk a v každém případě je to dost nepohodlný postup. Můžeme jej však dobře imitovat. Prostě jednotlivce v seznamu populace očíslujeme a pak použijeme "něco" co produkuje náhodná čísla a vybíráme ty jedince, jejichž číslo se s těmi náhodnými shoduje. Říká se tomu

prostý náhodný výběr

Jednoduchá však v tom není generace těch náhodných čísel. Kdysi se k tomu užívala taková podivná "kostka", mnohohran s deseti stejnými plochami, na každé z nich byla jedna z číslic od 0 do 9. Prý bylo obtížné vyrobit takovou "kostku", aby byla "poctivá", to je aby každá číslice měla stejnou pravděpodobnost, že "padne". Ještě do nedávna jsme používali tabulky náhodných čísel, dost tlusté knihy číselných skupin, o nichž nám matematici řekli, že v nich za takových a takových okolností nebyli schopni objevit žádnou pravidelnost. Dodnes jsou výtahy z těchto tabulek přetiskovány téměř v každé učebnici výzkumných metod. Jejich

správné používání rozhodně není nejzábavnější kratochvíle, ale někdy nám prostě nezbude nic jiného. Naštěstí dnes každá lepší kalkulačka a ovšem každý, i nejmenší osobní počítač umí produkovat náhodná (matematik by řekl "quasi-náhodná") čísla. Tenhle přístup má velikou výhodu: program produkuje náhodná čísla jenom v tom rozsahu, v jakém je potřebujeme. Řekneme počítači, jak je velká populace, třeba 300 a program pro nás vyprodukuje náhodná čísla jenom v rozsahu od 1 do 300. Tabulky náhodných čísel jsou nejméně pěticiferné. Pro naši velikost populace použijeme ovšem jen první nebo poslední tři sloupce číslic, ale i tak sedm z deseti nalezených nebudeme s to použít. Kalkulačka nebo počítač jsou mnohem efektivnější, a když si s tím nevíte rady, obraťte se na sousedova syna, a pokud by neměl takový program, většina těch chytrých holek a kluků, kteří vlastní třeba i ten nejmenší Sinclair, je schopna napsat takový program v Basicu za několik minut.

Dr. Watson:

Ale já nemám kalkulačku a všichni sousedi jsou bezdětní. Tak bych si to chtěl zjednodušit. Populace má 500 členů a já chci vzorek ve velikosti 100. Proč bych nemohl vzít jednoduše každou pátou osobu ze seznamu?

Tentokrát Dr. Watson promluvil pro změnu moudře. Technika, kterou navrhl se opravdu používá. Říká se jí systematický výběr. Nenechte se však zmást tím názvem; je to opět technika náhodného výběru.

Systematický výběr:

V systematickém výběru je do vzorku zahrnuta každá N-tá jednotka ze seznamu. Velikost kroku (N) dostaneme, když vydělíme velikost populace velikostí požadovaného vzorku. Důležité však je, aby první jedinec byl vybrán náhodně a teprve od tohoto výchozího bodu budeme vybírat každou N-tou jednotku.

Tento postup však nemůžeme použít, když jsou seznamy řazeny podle nějakého systematického schématu. Naše pohádka ilustruje něco, co se v praxi opravdu stává.

Pohádka pro odrostlejší děti 10.

O výběru, který byl příliš systematický

Bylo, nebylo, kdesi existovalo malé království, které se jmenovalo Org. Bylo to království, kde všechno bylo velice dobře zorganizováno, a přesto byl každý šťastný a spokojený. Každý, až na vojáky základní služby. Ti si stěžovali na plat, na stravu, na zacházení od představených, na všechno. a protože vše bylo dobře zorganizováno, vláda pozvala zahraničního odborníka, profesora P.I. Tomu, aby provedl výzkum postojů v armádě.

P.I. Toma přijel, zkonstruoval výborný dotazník a vyzkoušel jeho validitu. Protože to království bylo tak malé, že se tam ani počítač nevešel a místní knihovny neměly tabulku náhodných čísel, rozhodl se použít pro konstrukci vzorku techniku systematického výběru. Armáda toho malého království byla taky malá, důstojníci, poddůstojníci i mužstvo dohromady jen 12.000 osob. Profesor P.I. Toma odhadl, že vzorek 200 osob mu poskytne přijatelný interval spolehlivosti a zvolil tedy krok 60. Náhodně vybral prvého jedince. Byla to osoba č. 31 a pak vybíral každého dalšího šedesátého vojáka. Výsledky výzkumu byly prostě náramné. Ještě nikdo nikde nezkoumal tak spokojenou armádu. Každý byl šťastný v tom malém šťastném království - až do příštího jara, kdy začalo krvavé povstání vojáků základní služby.

Ale vy už víte, co se stalo: Prostě, v království Org vše bylo dobře organizováno. I seznamy členů armády byly uspořádány po četách, v každé četě nejdříve dva důstojníci, pak tři poddůstojníci, pak mužstvo základní služby a každá četa měla ne více, ne méně než 30 osob. a náš profesor měl smůlu, protože zvolený krok se shodoval přesně nejen s dvojnásobkem velikostí čety, ale také proto, že první náhodně vybraná osoba byl důstojník a tedy každá následující osoba musela být také důstojník. Poddůstojníci a vojáci základní služby nebyli zahrnuti do vzorku vůbec.

Nemysleme si, že takové zkreslení patří jen do absurdního světa pošetilých pohádek. Mnohé ze seznamů populací jsou systematicky uspořádány, kupř. žáci škol podle tříd, dělníci podle dílen atd. Někdy systém, podle kterého je seznam organizován, nemusí být na prvý pohled zřejmý. Kupř. byty na sídlištích ve velkých obytných budovách bývají identifikovány třícifernými čísly. Prvá číslice definuje podlaží, druhé dvě byt na podlaží. Protože půdorys se na každém podlaží opakuje, byty se stejnými posledními číslicemi budou mít obdobné vlastnosti, budou třeba větší či menší než byty ostatní, a to by opět při systematickém výběru mohlo produkovat zkreslení.

Podívejme se teď na jiný typ náhodného výběru, která by býval mohl zachránit profesora P.I. Tomu před zmíněnou blamáží:

Náhodný stratifikovaný výběr: Populace je rozdělena do skupin homogenních vzhledem k nějakému jasnému kritériu a jedinci jsou vybíráni do vzorku náhodně z těchto skupin.

Profesor Toma měl začít s třemi seznamy; se seznamem populace důstojníků, s jiným, zahrnujícím jen poddůstojníky, a konečně se seznamem vojáků základní služby. Z každé populace by pak byl vybrán náhodný vzorek, třeba technikou systematického výběru, a v našem malém království by k povstání třeba nedošlo. Ve skutečném světě, například při výzkumu studentů určité školy, bychom vybírali jedince zvlášť pro každý ročník. Při jiných výzkumech by populace mohla být stratifikována podle volebních obvodů, při výzkumu zaměstnanců továrny by mohl být výběr prováděn zvlášť mezi dělníky a zvlášť pro administrativu.

Stratifikovaný náhodný výběr má ještě jednu dodatečnou výhodu: snižuje velikost směrodatné chyby, a tedy i interval spolehlivosti. Třeba si ještě pamatujete, že chyba klesá s rostoucí velikostí vzorku a s přirůstající homogenností populace. Logika toho je zřejmá: když v populaci je pro kandidáta A 98% voličů a pro kandidáta B jen 2%, předpovědět, kdo vyhraje volby, je mnohem snadnější, než kdyby preference byle třeba 55% pro A a 45% pro B. Ve stratifikovaném výběru jsou vzorky podskupin zcela homogenní vzhledem k proměnné, podle které byly stratifikovány: ve skupině jsou jenom vojáci základní služby, nebo jenom posluchači druhého ročníku atd. Pro stratifikační proměnnou je tedy směrodatná chyba nulová a pro všechny jiné proměnné, které jsou s touto proměnnou asociovány, bude tato chyba podstatně menší.

A teď se podíváme na velmi zvláštní typ výběru, na vícestupňový náhodný výběr. Je to technika velice pracná, náročná a drahá, ale, jak hned uvidíme, velice důležitá a nenahraditelná.

Vícestupňový náhodný výběr

se provádí ve dvou nebo více krocích. Nejdříve jsou náhodně vybrána určitá přirozená seskupení, a pak teprve jsou náhodně vybíráni jedinci z oněch vybraných seskupení.

K čemu je to dobré? Pro ilustraci jednoho aspektu vás pozvu na výlet na jiný kontinent. Představte si, že bychom měli dělat výzkum na náhodném vzorku reprezentujícím dospělé obyvatelstvo Kanady. Kanada má něco přes dvacet milionů obyvatel, ale její plocha je téměř

10.000.000 čtverečných kilometrů. Řekněme, že velikost vzorku by byla 1.000 jedinců, a tak bychom teoreticky měli jednoho respondenta na deset tisíc čtverečných kilometrů. Ve skutečnosti by to bylo mnohem méně, obrovské rozlohy země jsou prázdné. Ale i tak jsou rozměry země obrovské a takové by byly i náklady. Při dané velikosti vzorku bychom měli nejmenší potíže s nejlidnatějšími provinciemi. V Quebecku bychom měli asi 290 respondentů, v Ontariu přibližně 350. Ale v Northwest Territories jednoho, nebo dva a ti by nás přišli pěkně draho. Pokud bychom neměli velké štěstí, museli bychom, abychom je zastihli, najmout hydroplán, helikoptéru nebo psí spřežení. Ale i v nejlidnatějších provinciích, a nebo i v prostorově malé zemi s tak vysokou hustotou obyvatelstva jako má Československo, rozptyl populace v prostoru podstatně zvyšuje náklady a nesmírně ztěžuje organizaci výzkumů. (Kupř. tazatelské týmy jsou organizovány a školeny lokálně; to snižuje cestovní náklady. Ale je jen omezený počet terénních center, které jsme schopni organizovat a financovat.) Tady je právě oblast uplatnění vícestupňového náhodného výběru. Můžeme postupovat třeba takto:

1. Nejdříve vybereme náhodně reprezentativní soubor okresů.
2. Pak v každém z vybraných okresů provedeme náhodný výběr obcí.
3. Ve velkých vybraných obcích zařadíme ještě další mezistupeň výběru: vybereme náhodně menší prostorové jednotky, třeba volební obvody.
4. Teprve pak vybíráme jedince.

Tímto způsobem obdržíme mnohem kompaktnější vzorek. Respondenti nejsou rozptýleni po celém teritoriu, ale jsou koncentrováni do zvládnutelného počtu regionů. Je-li takový výběr proveden správně, žádné závažné zkreslení reprezentativnosti nehrozí.

Nicméně existuje ještě jedna, dokonce důležitější doména použití tohoto výběru. Největším problémem pro použití pravděpodobnostního výběru v sociologii je fakt, že pro mnoho zajímavých populací žádný seznam neexistuje. Pro mnoho těchto situací je vícestupňový náhodný výběr jediným řešením. Řekněme, že bychom chtěli vytvořit pravděpodobnostní vzorek celé země a žádné spolehlivé seznamy obyvatelstva buď neexistují, nebo nejsou výzkumníkovi dostupné. To je mimochodem situace ve většině zemí světa.
Postup by byl shodný v prvých třech krocích s předchozí tabulkou, ale pak by následovaly dva další, logicky jednoduché, ale pracovně náročné kroky:

4. Ve vybraných malých obcích, nebo městských obvodech, je proveden soupis všech sídelních jednotek (bytů, rodinných domků).
5. Pak je vytvořen náhodný vzorek těchto jednotek.
6. Je vytvořen seznam osob žijících ve vybraných jednotkách a pak jsou opět náhodně vybráni jedinci (nebo obvykle jedinec) do vzorku.

Nejnáročnější je ovšem krok č.4. Představuje obsáhlou práci jak v přípravě, tak i v terénu; záznamy se obvykle opožďují za skutečností, nemusí rozlišovat mezi jednotkami, které jsou obydleny a těmi, které jsou používány pro jiné účely atd. Poslední krok je obvykle prováděn tazatelem přímo v terénu. Náhodnost musí být zaručena i při tomto kroku. Záznamový arch pro interview obsahuje instrukci, v jakém pořadí mají být členové domácnosti zaznamenáváni, a náhodně generované pořadové číslo osoby, která má být interviewována. Bez takové instrukce by tazatel vybral osobu, která je právě dosažitelná, aby se tak vyhnul nutnosti další návštěvy, nebo osobu, která je mu sympatická. Tak by byly kupř. podreprezentovány osoby, které během dne pracují mimo dům.

Někdy aplikace vícestupňového výběru nemusí být obtížná a je přitom velice užitečná. Chtěli bychom třeba studovat na celostátním vzorku mínění studentů dvou nejvyšších ročníků střední školy. Ústřední seznam středoškolských studentů asi neexistuje, ale existuje seznam všech středních škol a každá škola má seznam žáků, sestavený pravděpodobně podle ročníků. Výběr by mohl být prováděn třeba takto: Náhodně by byly vybrány okresy, pak vzorek škol v těchto okresech a jedinci do vzorku by byli náhodně vybráni ze seznamu žáků posledních dvou ročníků.

Před časem jsme zkoumali postoje starších osob k možnosti vstoupit do institucí pro staré občany (Disman & Disman, 1989). Naším cílem bylo sledovat vliv etnické kultury na tyto postoje; porovnávali jsme postoje Portugalců a Italů žijících v Torontu, ve věku 65 nebo starších, s postoji stejně starých Kanaďanů, jejichž mateřským jazykem je angličtina.

Vytvoření vzorku nebylo snadné. Osoby starší než 65 let představují 11% torontské populace, z těchto starších osob je jen 5% Italů a 1% Portugalců. (To znamená, že Portugalci ve věku 65 a více představují asi 0.11% z torontské populace.) Kdybychom tedy chtěli interviewovat

100 Italů a 100 Portugalců, museli bychom kontaktovat asi 100.000 domácností, a to je ovšem nemožné přinejmenším z finančních důvodů. Naštěstí jsme měli k dispozici seznamy osob pro daňové účely a tyto seznamy zahrnují prakticky všechny dospělé občany. Nadto tyto seznamy zahrnovaly také informaci o věku. Tato informace podstatně zúžila velikost vzorku pro vyhledávací fázi výzkumu. Ale i tak, abychom vyhledali vzorek 100 portugalských respondentů, museli bychom kontaktovat asi 10.000 domácností a i to by bylo nemožné.

Zůstala pro nás tedy otevřena jediná možnost: kontaktovat osoby ze seznamu, jejichž jména znějí italsky nebo portugalsky. Jistě, tato metoda má některé nevýhody. Kupř. portugalské jméno může mít britská manželka portugalského manžela, ale tyto případy byly vyloučeny v předběžném rozhovoru. Do vzorku nebyly zahrnuty osoby s etnicky netypickými jmény, italské nebo portugalské manželky mužů jiného etnického původu atd. Nicméně toto zkreslení - zejména vzhledem k silné tendenci obou národnostních skupin uzavírat sňatek uvnitř etnické skupiny (endogamy) nebylo příliš vážné. Ale i tak - zejména vzhledem k úmrtnosti mezi staršími osobami, vinou nepřesnosti záznamů, a vzhledem ke značné horizontální mobilitě - bylo nutno kontaktovat 652 portugalských adres, s výtěžkem 161 jmen respondentů, odpovídajících naší definici populace.

V tomto případě jméno jako kritérium pro výběr - doufejme - nezpůsobilo vážné zkreslení. Ale nemusí tomu tak být vždycky. Mezi americkými sociology koluje hezká historka, kterou uvedeme v naší pohádce č.11.

Pohádka pro odrostlejší děti 11.

O zrádném písmenu

Bylo před místními volbami v jednom velkém městě na východním pobřeží U.S.A. a skupina politiků si objednala výzkum, předpověď výsledků voleb. V té době mělo město dobrý seznam voličů řazený abecedně. Kartotéky zaplňovaly několik místností. Pro konstrukci vzorku byla použita technika vícestupňového náhodného výběru. Nejdříve byla vybrána náhodně místnost, pak kartotéční skříň a ze zásuvek této skříně byli vybráni technikou systematického výběru jedinci do vzorku.

Výzkum skončil neslavně: jako vítěze vyhlásil kandidáta, který skončil daleko vzadu v poli poražených. Prostě výzkumník se dopustil omylu, ale zejména měl smůlu. Náhodně vybral začátek písmena M, a tak se stalo, že voliči irského a skotského původu, jejichž jména velice často začínají na Mac a Mc, byli silně přereprezentováni. Hlasování ve volbách v USA a Kanadě velmi často sleduje etnickou linii. Není proto divu, že výzkum mylně předpověděl vítězství irského kandidáta. Tomuto zkreslení bylo snadné zabránit, kdyby byl stejný počet voličů vybrán z více kartotéčních skříní. Je ale také pravda, že kdyby bylo vybráno jiné písmeno, ke zkreslení by asi nedošlo.

5.4. Když koruna nepracuje

Zatím jsme viděli členy dvou rodin výběrových technik. Nejdůležitější jsou pravděpodobnostní techniky, založené na náhodném výběru. Jsou velice mocné, zajišťují, že budou dobře reprezentovány všechny známé i neznámé vlastnosti populace. Nadto jen u nich jsme schopni prostředky statistiky odhadnout, nakolik se vzorek liší od populace. Bohužel, zdaleka ne vždy jsme schopni tyto techniky použít. Někdy třeba proto, že pracnost a nákladnost těchto technik přesahuje rámec našich možností. Jindy proto, že neexistuje žádný seznam cílové populace. Nejčastější překážkou je však kombinace obou těchto důvodů. Speciální populace, o kterou se zajímáme, může být rozptýlena mezi celou populací a mít velice nízkou frekvenci. Teoreticky by bylo jistě možné vytvořit veliký vzorek celé populace a pak, po předběžných rozhovorech, vybrat jen ty jedince, kteří odpovídají definici naší cílové populace. Jak jsme si ilustrovali na příkladu výběru starých Portugalců, z hlediska nákladů by to bylo prostě nemožné. My jsme měli štěstí, byli jsme schopni zimprovizovat seznam populace, ale to se stává spíše výjimečně.

Jako prvou techniku tvorby vzorku jsme v této kapitole diskutovali kvótní výběr. Reprezentuje druhou skupinu výběrových technik, které nejsou založeny na teorii pravděpodobnosti, ale na logickém úsudku. Kvótní výběr je pravděpodobně nejspolehlivější mezi těmito technikami, ale opět ne vždy je možno jej použít. Může být aplikován jen tehdy, když máme dostatečnou znalost o populaci, abychom její strukturu mohli imitovat ve struktuře vzorku. Do téže skupiny patří účelový výběr, někdy i anketa, a bývá sem zařazována i technika sněhové koule (snowball sampling).

Účelový výběr
je založen pouze na úsudku výzkumníka o tom, co by mělo být pozorováno a o tom, co je možné pozorovat.

Jak vidíte, není to příliš vědecký přístup, ale velice často jediný, který nám zbývá. Je používán i profesionálními agenturami, které provádějí za úplatu výzkum trhu. Řekněme, že byste při sobotním nákupu v Bílé labuti byli osloveni mladým mužem a dotazováni na to, co si myslíte o určité skupině výrobků. Na jakou populaci se výsledky takového výzkumu vztahují?

Dr.Watson

Ale to je přece jednoduché! Na lidi, kteří nakupují v obchodních domech!

Jako obvykle, Dr. Watson je příliš optimistický. Takto konstruovaný vzorek by reprezentoval přinejlepším populaci osob, které nakupují v Bílé labuti v sobotu dopoledne a právě v této roční době. A kdybychom měli být opravdu přesní, museli bychom ještě dodat, že se závěry vztahují jen na ty jedince z takto definované populace, kteří jsou ochotni odpovídat na otázky daného typu. Není to příliš široká a dobře definovaná populace, ale pro účely výzkumu trhu by mohly být takto získané informace určitě užitečné.

Účelový výběr nám téměř nikdy neumožní nějakou opravdu širokou generalizaci našich závěrů, ale to neznamená, že tyto závěry nejsou užitečné. Jen nesmíme předstírat jiným, a především ne sobě, že tyto závěry platí pro každého jedince ve vesmíru.

Při použití účelového výběru musí výzkumník jasně, přesně a otevřeně definovat populaci, kterou jeho vzorek **opravdu** reprezentuje.

Užití účelového výběru je pro některé populace jediným řešením. To platí kupř. pro etnické minority; snad v žádné zemi neexistují spolehlivé a vyčerpávající seznamy takových skupin. Pak nezbývá nic jiného, než použít jako výchozí bod seznamy členů etnických organizací. Takový vzorek jistě nebude reprezentovat ty příslušníky etnické skupiny, kteří nejsou v žádné z takových skupin organizováni. Je pak na výzkumníkovi, aby posoudil, s použitím znalosti skupiny, jak dalece jsou jeho závěry zobecnitelné. Kupř. Pejovič (1990) zkoumal vzdělávací aspirace středoškolských studentů chorvatského původu, žijících v Torontu. Jeho závěry jsou velice zajímavé a závažné. Zdá se, že pro tuto skupinu neplatí obvyklé socioekonomické determinanty aspirací, které americká sociologie má tendenci považovat za univerzální. Pejovič užil techniku účelového výběru. Východiskem bylo členstvo různých chorvatských kulturních a sociálních organizací, účastníci různých chorvatských společenských akcí atd. Pejovič nikde nepředstírá, že jeho závěry platí mimo jeho vzorek. Nicméně síla zjištěných souvislostí a známá fakta o kulturní a socioekonomické homogenitě této etnické skupiny naznačují, že je silně pravděpodobné, že podobné výsledky bychom mohli dostat i pro většinu jiných mladých Chorvatů, žijících ve velkých kanadských městech. **Důkaz** pro to by ovšem mohl být získán jen opakováním výzkumu na vzorcích vytvořených z dalších populací. Tedy i technicky vzato nereprezentativní vzorek může někdy poskytnout hodnotné výsledky. Ne však vždycky a ne automaticky a musíme si být vědomi toho, že je jen náhražkou za pravděpodobnostní výběr.

Některé techniky vytváření účelového vzorku jsou velice problematické. Bohužel, stále je hojně používána anketa.

V anketě je výběr jedinců založen na rozhodnutí respondenta zodpovědět otázky uveřejněné v masových sdělovacích prostředcích.

Definovat populaci, ke které se nálezy ankety vztahují, je skutečně nemožné. Nejsou to čtenáři určitých novin nebo časopisu. To by bylo ještě dobré. Vzorek se však liší od celé populace právě tím, že to jsou ti, kteří zodpověděli anketu. Maximálně můžeme říci, že lidé ve vzorku jsou více motivováni, než ostatní čtenáři, a to je velice slabá definice. Ani velikost vzorku nepomůže. Správně konstatuje Zich (1976, str. 207) že anketa Rudého práva, která získala vzorek větší než 110 tisíc není reprezentativní, i když v základních demografických ukazatelích se dosti shodovala se strukturou celé dospělé populace. Problém je v samovýběru respondentů. Ale to už známe z našeho příkladu vojáků, filmu a postojů k U.S.A. Poznávací hodnota ankety je podle mého názoru pod hodnotou dobře a zodpovědně napsaného fejetonu.

A konečně tu máme techniku "snowball sampling", techniku sněhové koule. Podle mého názoru tato technika vůbec do této kapitoly nepatří. Je to technika identifikace populace, a ne vytvoření reprezentativního vzorku. Ale všechny učebnice, které znám, ji zařazují mezi výběrové techniky, a tedy i my sledujeme toto schéma. Ale posuďte to sami.

> "Snowball Technique" spočívá na výběru jedinců, při kterém nás nějaký původní informátor vede k jiným členům naší cílové skupiny.

Nejlépe si to ukážeme na jednoduchém příkladu. Třeba bychom chtěli studovat mocenskou strukturu v malé obci. Identifikovat oficiální vlivné osoby, jejichž pozice je formálně definována by nebylo těžké. U nás by to byl třeba starosta, v nedávné minulosti straničtí funkcionáři, SNB atd. Ale vliv v obci mohou mít osoby, jejichž vliv není definován funkcí, a tato část souboru vlivných osob se liší podle místních okolností. V některé obci může být vlivnou osobou ředitel školy či továrny, v jiné obci mohou být osoby v takových funkcí bez vlivu a významný vliv na rozhodování může mít kněz, nebo vlivný a mocný rodák, který v obci již dlouhá léta nežije. To vše je pro výzkumníka, který přichází z venku, neviditelné. Tady je na místě uplatnit výběr technikou sněhové koule.

Výzkumník začne rozhovorem s jasně definovanou osobou, třeba starostou. V tomto rozhovoru požádá respondenta, aby jmenoval další vlivné osoby. Ty jsou pak interviewovány a každá z nich dostane i stejnou otázku o vlivných lidech. Po určitém počtu rozhovorů se již jména nových vlivných osob neobjevují. Výzkumník může prohlásit, že vzorek je "teoreticky nasycen". Populace vlivných osob v obci byla jasně identifikována a náš vzorek je totožný s touto populací.

Technika sněhové koule, kde jména dalších osob se v řetězci rozhovorů "nabalují" jako sněhová koule (taková, kterou je znázorňována lavina v kreslených vtipech) je

nenahraditelným nástrojem pro zkoumání populací, které existovaly jen dočasně: účastníci určitých demonstrací, svědkové katastrofy nebo jiné řídké události atd. Zde většinou teoretické nasycenosti vzorku nedosáhneme a aplikace této techniky má opravdu charakter konstrukce účelového vzorku.

Termín "teoretická nasycenost" byl uveden Glaserem a Strausem (1967) v souvislosti s jejich konceptem "grounded theory", snad nejdůležitějším epistemologickým nástrojem pro kvalitativní výzkum. Technika sněhové koule hraje pod jménem "teoretický výběr" velice důležitou úlohu. Má zde do jisté míry funkci ověřování validity. Ale k tomu se ještě vrátíme s celou řadou podrobností. Doufám, že to bude docela zajímavé.

5.5. Koruna přece jen není všechno

Techniky náhodného výběru opravdu produkují nejlepší možnou reprezentaci populace. Jenomže je reprezentativní pouze za předpokladu, že všichni vybraní jedinci se opravdu na výzkumu zúčastnili, to jest, že například zodpověděli naše otázky. Výpočet směrodatné chyby je plně založen na tomto očekávání. V příští kapitole uvidíme, že je to příliš optimistický předpoklad. V současné době procento osob, které odmítly tazatele nebo nevrátily dotazník, téměř všude roste. U dotazníku návratnost často nedosáhne ani padesáti procent.

Dr.Watson

Ale to přece vůbec není problém. Já vím, že nám v jedné z dalších kapitol řeknete, že dotazník je jednou z nejlevnějších technik sběru dat. Tak když potřebuji ve vzorku 300 jedinců, prostě rozešlu 900 dotazníků a tak dostanu vzorek i větší, než skutečně potřebuji.

Jistě už víte, proč by tento recept nefungoval: populace, která odpověděla, není totožná s tou, která odmítla odpovědět. Liší se v něčem, co bylo důvodem pro toto rozhodnutí, a pravděpodobně ono "něco" je silně spojeno s problémy, na které je výzkum zaměřen. Obvykle jsme o těchto důvodech schopni jenom spekulovat. Obávám se, že tu musím uvést nový typ nepříjemné redukce informací:

> Redukce negativním samovýběrem vzniká tehdy, když část jedinců, vybraných do vzorku, odmítla na výzkumu participovat. Tento typ redukce může vážně ohrozit reprezentativnost vzorku.

Toto je vážný problém. Tak vážný, že před několika lety byl ústředním tématem výročního zasedání Americké statistické společnosti. Vidíte, na začátku této kapitoly jsme si pochvalovali, že redukce populace na vzorek je logicky, technicky a metodologicky dobře propracovanou operací, kde riziko zkreslení je menší, než v jiných výzkumných operacích. Je to stále pravda, ale přece i zde máme zranitelné místo. Neznáme žádný univerzální lék na tento neduh. Jediné řešení je usilovat o co nejvyšší návratnost. U některých technik sběru informací je to snadnější, u některých je to téměř nemožné. Ale tohle už patří do příští kapitoly.

Řešení úkolů z kapitoly 5.

Cvičení 5.1.

Jistěže ne. Lidé, kteří neměli telefon (a těch bylo v roce 1936 velice mnoho), ti kteří nevlastnili auto, nebyli členy organizací, tedy lidé náležející do nižších socioekonomických vrstev byli ze vzorku opravdu vyloučeni.

Cvičení 5.2.

Tohle nebyla poctivá otázka. Kvótní výběr může být aplikován jen na populaci, jejíž vlastnosti relativně dobře známe. V našem případě bychom mohli nanejvýše navrhnout něco o taxikářích, vrátných v hotelech a prý i příslušnících bezpečnostních orgánů, ale rozhodně by to nebylo dost pro konstrukci kvótního vzorku.

Kapitola 8.

Já ti to spočítám
aneb
Statistika pro úplně beznadějné případy.

Důrazné varování!
Statistik může číst tuto kapitolu jen na vlastní nebezpečí!

Opravdu, tohle není statistika. To dokonce ani není úvod do statistiky. Je to spíš pokus vysvětlit, jaký **logický význam pro interpretaci dat** mají operace, kterými nám statistika pomáhá. Doufejme, že tahle kapitola ukáže, že ta hrozivá statistika je nejen užitečná, ale i docela zajímavá. Třeba alespoň některým z nás pomůže překonat strach a přiměje nás začít se zabývat statistikou trochu vážněji.

8.1. Kdo je v obálce?

Už jsme si řekli několikrát, že kvantitativní výzkum není nic jiného než testování hypotéz. Také už víme, že pracovní hypotézy jsou v podstatě předpovědí, jaká by byla souvislost mezi proměnnými, kdyby naše hypotézy byly pravdivé. Ale co to slovo **"souvislost"** vlastně znamená? Podívejme se na to na chvíli z celkem zajímavé perspektivy hazardního hráče.

Představte si, že máte před sebou zapečetěnou obálku a víte jenom, že je v ní vyplněný dotazník z výzkumu na celostátním vzorku dospělého obyvatelstva. Váš úkol je uhodnout, jaké je pohlaví respondenta, jehož dotazník je v obálce. Jaká je pravděpodobnost, že uhodnete správně?

To byla lehká otázka. V populaci je přibližně stejné procento mužů jako žen. Máme tedy přibližně pravděpodobnost jedna ku jedné, že budeme mít pravdu. Totéž, abychom udělali

lepší dojem na kolegy, můžeme vyjádřit v odborném žargonu: pravděpodobnost správného odhadu je 50%. Nebo ještě jinak: $p = 0.5$. Změňme teď poněkud podmínky naší hry. Představme si, že v obálce je vyříznuté okénko, kterým vidíme odpověď na následující otázku:

Užíváte někdy rtěnku?

ANO
NE

Bude-li odpověď, ANO i nepříliš chytrý hazardní hráč bude hádat, že dotazník zodpovídala žena. Může se stále ještě mýlit - i muži mohou používat rtěnku. Když by odpověď byla NE, bylo by lépe hádat muže. Ovšem, i zde se můžeme mýlit; mnoho a mnoho žen nepoužívá rtěnku. Ale rozhodně pravděpodobnost správného odhadu je mnohem vyšší, než byla předtím, než jsme získali informaci o používání rtěnky. Informace o rtěnce zvýšila pravděpodobnost správného odhadu respondentova pohlaví. Můžeme tedy říci, že mezi proměnnými "pohlaví" a "používání rtěnky" existuje souvislost.

Souvislost může být definována jako přírůstek v pravděpodobnosti správného odhadu jedné proměnné, za který vděčíme naší znalosti o jiné proměnné.

Můžeme-li odhadnout stav jedné proměnné ze stavu jiné proměnné bez jakéhokoliv omylu, mluvíme pak o **perfektní** nebo **deterministické** souvislosti. Ale tím se my v sociálních vědách vůbec nemusíme zabývat. Proměnná perfektně souvisí jenom sama se sebou.

Dr.Watson:

Počkejte, to přece nemusí být pravda. Souvislost mezi proměnnými "pohlaví" a "těhotenství" je perfektní. Když odpověď na otázku o těhotenství je pozitivní, vím bez jakékoliv možnosti omylu, že respondent byla žena?
My: Ne tak rychle, drahý doktore. Co když odpověď byla negativní?
Dr.Watson zahanbeně mlčí.

A tady jsme zase nazpět u naší staré bolesti, velikosti přirozených systémů v sociálních vědách, kde všechno souvisí se vším. To je také důvodem, proč nenajdeme nikdy deterministické spojení mezi proměnnými.

Cvičení 8.1.

Vysvětlete prosím doktoru Watsonovi, proč nemůžeme nalézt perfektní souvislost. Pokud máte nějaké potíže s tímhle vysvětlením, podívejte se znovu na první kapitolu.

Souvislost může mít mnoho různých forem. Podívejme se třeba na tabulku 8.1. Sloupce tabulky reprezentují známky z matematiky, řádky známky z deskriptivy. A je nejlepší známka, D nejhorší. Čísla v polích tabulky reprezentují **relativní sloupcové četnosti**. To znamená, že kupř. nejvyšší pole v prvém sloupci odpovídá tomu procentu ze všech studentů se známkou A v matematice (50 %), kteří také dostali A z deskriptivy. Součet čísel v každém sloupci reprezentuje 100%. Řádek N pak obsahuje informaci o počtu pozorování, obsažených v každém sloupci.

Říká nám tato tabulka něco o souvislosti mezi známkou z deskriptivy a matematiky? Je tu nějaká informace, která by pomohla hazardnímu hráči volit optimální strategii jak hádat, jakou známku dostal určitý student z deskriptivy, když zná jeho známku z matematiky? Řekněme, že student dostal z matematiky známku A. Náš hazardní hráč by se podíval do prvého sloupce tabulky. Zjistil by, že mezi studenty, kteří dostali A z matematiky je nejčastější známka z deskriptivy také A. Vsadil by se tedy, že známka z deskriptivy je A. Měl by sice stále právě padesátiprocentní pravděpodobnost, že prohraje, ale každá jiná strategie by mu nabízela menší naději výhry. (Kdyby to nebyl hráč, ale sebevrah - nechť pan Nezval promine - měl by **jistotu**, že prohraje.) Kdyby student dostal D z matematiky, to by se to teprve hádalo: hráč by navrhl, že student dostal také D z deskriptivy a měl by pěkně vysokou naději na výhru: celých 80%. Podobně by mohl postupovat pro každou jinou hodnotu známky z matematiky.

Tabulka 8.1.

DESKRIPTIVA MATEMATIKA

	A	B	C	D
A	50%	35%	10%	0%
B	45%	55%	25%	10%
C	5%	8%	55%	10%
D	0%	2%	10%	80%
N	100% 150	100% 360	100% 400	100% 50

Kupení vysokých hodnot na diagonále naší fiktivní tabulky nám dosti důrazně naznačuje, že existuje nějaká souvislost mezi známkami z matematiky a deskriptivy. Zdá se, že studenti s dobrou známkou z matematiky mají také dobrou známku z deskriptivy a studenti se špatnou známkou mají opět špatnou známku.Kdyby všechna pozorování byla nahromaděna jenom na diagonále (v tmavých polích tabulky) a ostatní pole by obsahovala jenom nuly, tabulka by vyjadřovala perfektní souvislost. Hráč by přestal být hráčem, už by nehádal, už by věděl. Znalost o stavu jedné proměnné by mu poskytla úplnou informaci o stavu druhé proměnné.

Cvičení 8.2.

Co by znamenalo, kdyby všechna pozorování byla nakupena na druhé diagonále tabulky, vedoucí z levého dolního rohu do horního pravého rohu?

Tohle nebylo příliš těžké, že ano? Viděli jsme, v jaké formě se může souvislost projevit v jedné skupině proměnných. Ale souvislost může mít ještě jiné formy. Zkusme si ještě jiný příklad.

A teď se podívejme na jinou tabulku. Sloupce v této tabulce představují proměnnou X, která má kategorie A, B, C, a D. Řádky reprezentují proměnnou Y. Ta má kategorie J, K, L a M. Všechna čísla jsou jenom v jejích tmavých polích. Je nějaká souvislost mezi proměnnými v tabulce 8.2.?

Tabulka 8.2.

Proměnná X

Proměnná Y

	A	B	C	D
J				
K				
L				
M				

Dr. Watson se horlivě hlásí:
To je jednoduché. Tady není žádná pořádná souvislost. Na diagonále není doslovně nic. Ani na jedné z nich!

Dr. Watson byl stvořen tak, aby se mýlil, tedy za jeho omyly můžeme my. Tenhle je docela typický. Tahle souvislost má prostě jinou formu. Ale řešení je opět docela prosté. Řekněme, že proměnná X v tabulce 8.2. představuje nějakou územní dimenzi, že kategorie A až D

reprezentují volební obvody. Proměnná Y reprezentuje politické strany, pro které by respondent mohl hlasovat. A teď se podívejme znovu na distribuci dat v tabulce. Vidíme, že **všichni** respondenti z obvodu A preferují politickou stranu K, že všichni z obvodu B hlasují pro stranu M atd. V každém sloupci jsou pozorování nahromaděna do jediného pole a pozice tohoto pole je pro každý sloupec jiná, jedinečná pouze pro tento sloupec. Když známe hodnotu X, odhadneme hodnotu Y bez jakéhokoliv omylu. Tedy naše fiktivní data představují jinou formu perfektní, deterministické souvislosti.

Ve skutečnosti by taková souvislost měla jinou formu. V jakékoliv společnosti s minimální úrovní demokracie by - doufejme - prázdná pole neexistovala. Pokud bychom zjistili, že proporce respondentů hlasujících pro určitou politickou stranu v určitém obvodu se podstatně liší od proporcí v jiných obvodech, pak by se zvýšila pravděpodobnost správného odhadu volby politických stran na základě znalosti volebních obvodů. Celá řada důležitých statistických operací je v podstatě **založena na srovnání nalezené distribuce pozorování do polí tabulky s takovou distribucí, jakou bychom obdrželi, kdyby byla pozorování zařazena do polí tabulky náhodně.**

A teď se podívejme na ještě jinou formu, ve které může být souvislost vyjádřena. Chtěli bychom řešit následující důležitý problém: kdo mezi studenty sociologie konzumuje více piva, muži nebo ženy? Následující čísla se týkají **průměrného** počtu půllitrů vypitých během jednoho týdne:

Tabulka 8.3.

	Aritmetický průměr:
muži	8
ženy	2

Můžeme na základě těchto dat uzavřít, že existuje souvislost mezi proměnnými "pohlaví" a "spotřeba piva"? Zdravý rozum a Dr. Watson navrhují, že ano: muži v našem vzorku pijí pivo čtyřikrát více než ženy. Zdá se, že opravdu existuje souvislost mezi proměnnými "pohlaví" a "spotřeba piva". Ale taková data jako je aritmetický průměr nebo údaje vyjádřené v procentech představují velice podstatnou **redukci informace**. Zamyslete se třeba nad následující pohádku:

Pohádka pro odrostlejší děti 19.

O pohřešovaném kuřeti

Tuhle historku jste už asi slyšeli. Nevím, jaký je její původ, ale vypráví se po univerzitách a výzkumných ústavech celého světa. Je to v podstatě citát z výzkumné zprávy:

"Po aplikaci preparátu B se 33.3% kuřat uzdravilo, 33.3% uhynulo a o zbývajících 33.3% nejsme schopni poskytnout uspokojující informaci. Dosud se nám nepodařilo to třetí kuře chytit."

Morálka téhle pohádky je pro náš problém docela jasná: více bychom věřili průměru, který by byl vypočítán na vzorku 500 pozorování, než průměru vypočítaném pro vzorek pěti jedinců. Vzpomínáte, co jsme si řekli v kapitole 4. o intervalu spolehlivosti?

To ale ještě není všechno. Aritmetický průměr, stejně jako jiné podobné reprezentace středních hodnot redukuje informaci o mnoha jedincích do jednoho jediného údaje, a to je pěkně silná redukce, při které můžeme ztratit důležitý kus informace: Studujeme opět konzumaci piva ve dvou populacích. Pro obě populace jsme obdrželi zcela shodný průměr: 8 piv za týden. Můžeme tedy navrhnout, že jsou obě populace vzhledem ke konzumaci piva shodné? Pro spolehlivý závěr potřebujeme vědět, jak dobře průměr popisuje původní data. Uveďme si dva extrémní příklady, ilustrující původní data, ze kterých byl průměr vypočítán. Abychom ušetřili místo, předstírejme, že oba vzorky sestávaly jen z pěti jedinců:

Tabulka 8.3.

JEDINEC	Populace A: počet piv:	Populace B: počet piv
Jedinec 1.	8	0
Jedinec 2.	8	0
Jedinec 3.	8	0
Jedinec 4.	8	0
Jedinec 5.	8	40
Součet: Aritmetický průměr:	40 8	40 8

Je zřejmé, že průměr 8 reprezentuje skupinu A perfektně. Ale skupina B, to je docela jiná záležitost. To je vlastně skupina abstinentů, do které se vloudil jediný pivní hrdina, který nese obtížné břemeno: udržet průměrnou konzumaci piva na úrovni srovnatelné se skupinou A.

Je nesporné, že rozdíl mezi dvěma průměry signalizuje přítomnost souvislosti mezi proměnnou, podle které byli jedinci rozděleni do dvou subpopulací, a proměnnou popsanou jako průměr. Problém je jenom v tom, jak zjistit, že ten rozdíl mezi dvěma průměry je dostatečně významný. Teď už víme, že nestačí vzít v úvahu jen velikost vzorku, ale i to, jak je populace homogenní. Za chvíli se seznámíme s konceptem směrodatné chyby, která měří homogennost populace, ale hlavně, její diskuse nám umožní pochopit jiný důležitý koncept: statistickou významnost.

Ale ještě, než se podíváme, jak se souvislost opravdu měří, podívejme se, proč může mít souvislost tak mnoho různých tváří.

8.2. Statistika je třídní...

Vlastně třídní není ani tak statistika, ale proměnné, které můžeme, a jak uvidíte, musíme klasifikovat do několika skupin, které jsou vzájemně v hierarchickém vztahu. Je to důležité proto, že pro každou tu třídu proměnných můžeme použít jenom určitý soubor statistických operací. Skutečný statistik by vám předložil poněkud složitější třídění znaků a probral by i jiné principy pro klasifikaci proměnných, ale pro naši diskusi nám postačí podívat se na tři základní rodiny: **nominální, pořadové a intervalové proměnné.** (Zatajil jsem vám ještě jednu skupinu proměnných: alternativní znaky. Schoval jsem si to jako příjemné překvapení. Tak se, prosím, tvařte potěšeně, až je uvedu ve výkladu regresní analýzy.)

Nominální proměnné:

Říká se jim také kvalitativní proměnné. Jejich kategorie jsou pouhá jména a nedává mnoho smyslu se ptát, zda určitá kategorie je vyšší nebo nižší než jiná. Příkladem nominální proměnné je třeba respondentovo pohlaví, jeho barva vlasů, rodiště. To jsou proletáři mezi proměnnými. Řadu statistických operací, které můžeme používat pro ordinální a intervalové proměnné, nemůžeme zde uplatnit.

Pořadové proměnné

Říká se jim také ordinální proměnné. U těchto proměnných mohou být jejich kategorie seřazeny do nějaké hierarchie. Můžeme se smysluplně ptát, zda sledovaná vlastnost je u určitého jedince vyšší (nižší, silnější, lepší atd.) než u jiného respondenta. Nevíme však, o kolik je větší. Víme kupř., že stříbrná medaile je lepší než bronzová, ale ne tak dobrá, jako zlatá. Otázka, kolikrát je stříbrná medaile lepší než bronzová, však nedává smysl.

Intervalové proměnné

Ty mají takové kategorie, že nejen dává smysl se ptát, zda určitá kategorie je vyšší než jiná, ale také otázka, kolikrát je vyšší, je zde smysluplná. Příjem, věk, počet dětí jsou typickými ukázkami tohoto typu proměnných. Intervalové proměnné jsou aristokracií mezi ostatními proměnnými. Statistika s nimi může provádět taková kouzla, která nejsou dovolena pro nižší úrovně měření. Bohužel, intervalových proměnných není ve světě sociálního výzkumu mnoho.

Vidíme tedy, že by nemělo být obvykle příliš obtížné rozhodnout, do které skupiny určitá proměnná náleží. Stačí na ni aplikovat obě zmíněné kritické otázky a zamyslet se, zda jejich aplikace dává nějaký rozumný smysl.Následující shrnutí by nám mělo tento proces ulehčit.

Kritické otázky:

(A) Je určitá kategorie proměnné větší (menší) než jiná kategorie?
(B) Kolikrát je větší (menší)?

Jsou tyto otázky smysluplné?

A	B	
ne	ne	nominální proměnná
ano	ne	pořadová proměnná
ano	ano	intervalová proměnná

A opět slyšíme dr. Watsona brumlat někde v pozadí, k čemu je to všechno vůbec dobré. Důvod, proč potřebujeme vědět, do jaké skupiny určitá proměnná patří, je opravdu vážný: pro každou ze tří skupin proměnných můžeme použít jen určitý soubor statistických operací. Jako máme nominální, pořadové a intervalové proměnné, máme také nominální, pořadové a intervalové statistické operace. Jenže ty mají zajímavou hierarchii:

Nominální statistické operace nedovedou mnoho z toho, co dovedou operace vyššího řádu. Ale mají jednu příjemnou vlastnost: můžeme je aplikovat na nominální, právě tak jako na pořadové nebo intervalové proměnné.

Pořadové operace dokáží více než nominální, ale zdaleka ne tolik, co intervalové. Můžeme je aplikovat jen na ordinální a intervalové proměnné, ne však na nominální.

Intervalové statistické operace dokáží daleko více, než obě předchozí. Můžeme je však aplikovat výhradně jen na proměnné intervalového charakteru.

A zde je tato hierarchie vyjádřena v tabulce:

Proměnné	Nominální operace:	Pořadové operace:	Intervalové operace:
nominální	ANO	NE	NE
pořadové	ANO	ANO	NE
intervalové	ANO	ANO	ANO

Smysl našeho výkladu snad pochopíme lépe na následujícím příkladu, který nepatří do oblasti měření souvislosti mezi znaky, ale do oblasti popisné statistiky. Často je pro nás výhodné vyjádřit informaci o vzorku nebo o celé populaci v co nejjednodušší formě. Chceme kupř. říci něco o počtu dětí v rodině v Praze. Publikovat seznam všech rodin s počtem dětí by poskytlo velmi úplnou informaci, ale bylo by to dosti nepohodlné, nepřehledné, a z mnoha důvodů i prakticky nemožné. Proto se obvykle spokojíme s informací o průměrném počtu dětí. **Aritmetický průměr** je **intervalový popis střední hodnoty**. Můžeme jej tedy použít jenom pro popis intervalových dat, jako počet dětí, příjem, věk apod. Ale zjistit, jaká je průměrná barva očí studentů sociologie by byl z hlediska statistiky docela absurdní úkol. Pro proměnné na různé úrovni měření používáme odpovídající indikátory centrální tendence:

intervalová data	**aritmetický průměr**
pořadová data	**medián**
nominální data	**modus**

Aritmetický průměr, ten umíme všichni vypočítat: prostě sečteme pozorované hodnoty a vydělíme je počtem sledovaných jedinců.

Medián je ta hodnota, která je právě v prostředku všech pozorování, která jsme seřadili podle jejich velikosti. Seřadíme třeba děti ve třídě podle velikosti a velikost dítěte, které je právě uprostřed řady, reprezentuje medián.

Modus je prostě kategorie s nejvyšší četností. Zjistíme-li třeba, že studenty mají nejčastěji modré oči, "modrá" bude modus.

A teď se podívejme, zda opravdu platí to, co jsme si řekli o aplikovatelnosti různého typu statistik. Nominální měřítko, modus, by měl být a opravdu je aplikovatelný samozřejmě na nominální, ale i na pořadová data. Ale je aplikovatelný kupř. i jako charakteristika intervalové proměnné, jako kupř. počet dětí v rodině. Mohli bychom kupř. roztřídit rodiny v našem vzorku do kategorií podle počtu dětí. Kategorie, ve které jsme nalezli nejvyšší počet pozorování, t.zv. modální kategorie, může být pak použita jako charakteristika dané populace.

Někdy může být dokonce výhodné použít statistiky nižší úrovně. Jsou totiž méně citlivé k extrémním hodnotám. Podívejte se znovu na tabulku 8.4. a sledujete data pro populaci B. Aritmetický průměr už známe, ale 8 není právě přesvědčivou reprezentací této populace. Jaký bude medián? Jedinec číslo tři je právě v prostředku, hodnota sledované proměnné se rovná nule. To je, alespoň intuitivně, lepší reprezentace vzorku. Mezi odpověďmi na naši otázku se nejčastěji objevuje nula. Tedy modus se rovná nule. Prostě, medián a modus nebyly ovlivněny atypicky vysokou hodnotou odpovědi jedince číslo 5. Jistě jste si všimli, že pro zcela homogenní populaci a ze stejné tabulky aritmetický průměr, medián a modus mají stejnou hodnotu: osm.

Nejčastěji je však výhodnější zvolit "nejvyšší" typ statistické operace z těch, které smíme použít: tyto operace prostě dovedou mnohem více, než operace "nižší". Už víme, že střední hodnoty charakterizují vzorek tím lépe, čím je tento vzorek homogennější. Pracujeme-li s intervalovými proměnnými, můžeme popsat homogenitu vzorku docela chytrým způsobem. Nabízí se nám tu dva koncepty: **rozptyl** (variabilita) a jeho mnohem rafinovanější příbuzná, **směrodatná odchylka**. Rozptyl nám poskytne informaci, jak se pozorování v průměru liší od průměru. Ale...

Dr. Watson nás přerušuje:
Já už vím, jako to udělat: Nejdřív vypočítám aritmetický průměr, pro každého jedince odečtu pozorovanou hodnotu od toho průměru. Pak sečtu všechny ty odchylky dohromady a výsledek vydělím a dostanu...

A teď musíme přerušit my: "Nulu, pokaždé nulu, doktore!" Když by průměr vypočítán správně, součet negativních odchylek musí být přesně stejně veliký, jako součet pozitivních odchylek. Ale s výjimkou jediného kroku měl dr. Watson pravdu. Ukažme si teď, jak se to skutečně dělá. Protože jsem vám slíbil, že nebudu používat (skoro) žádné vzorečky, budeme si prostě povídat, jak to uděláme:

Návod k výpočtu směrodatné odchylky
(speciálně pro Dr. Watsona)

Co uděláme:	**Co to znamená**
1. Pozorovanou hodnotu pro každého jedince odečteme od vypočítaného průměru.	Vypočítali jsme soubor odchylek pro celý vzorek.
2. Odchylku vypočítanou pro každého jedince umocníme.	Negativní číslo násobené samo sebou ním dá vždy pozitivní hodnotu. Tím překonáme problém, na který narazil dr. Watson. Součet neumocněných odchylek by nám musel pokaždé dát nulu.
3. Umocněné odchylky sečteme.	Součet představuje souhmnou velikost (umocněných) odchylek.

4. Součet vydělíme počtem jedinců ve vzorku.	Často potřebujeme porovnat homogenitu souborů různé velikosti. Abychom kontrolovali rozdíly ve velikosti vzorků, vypočítáme, kolik z celkové sumy čtvercových odchylek připadá v průměru na každého jedince. Výsledek tohoto kroku je **rozptyl**, neboli **variabilita** sledovaného souboru.
5. Výsledek dělení odmocníme.	Rozptyl není měřen v jednotkách, ve kterých bylo pozorování prováděno, ale v mocninách těchto jednotek. Původně jsme pro každého jedince odchylky od průměru umocnili. Nyní kumulovaný průměrný výsledek opět odmocníme a dostaneme **směrodatnou odchylku.**

Cvičení 8.3.

Zkuste vypočítat rozptyl a směrodatnou odchylku pro populaci A a B z naší tabulky 8.4.

Směrodatná odchylka je svým způsobem magické číslo. Většina jevů v přírodě (bohužel ne tak docela automaticky ve společenských vědách) má tak zvané **normální rozložení**: Na stromě je nejméně hodně malých lístků. S přibývající velikostí stromových lístků jejich frekvence přirůstá a dosáhne maxima u listů střední velikosti. Když velikost lístků překročí průměrnou hodnotu, jejich četnost ubývá a opět, podobně jako tomu bylo s nejmenšími lístky, nejméně bude opět těch největších stromových listů. Podobnou distribuci - i když ne tak soustavně - objevíme i u řady sociálních jevů, jako výše příjmu, počet dětí v rodině, léta školního vzdělání atd. Můžeme si to snadno graficky znázornit. Na vodorovnou osu naneseme hodnoty, které může studovaná proměnná nabývat. Na kolmou osu pak naneseme množství případů (pozorování, jedinců) kteří mají danou hodnotu proměnné. Pak spojíme nalezené průsečíky křivkou. Máme-li hodně pozorování, dostaneme ladnou křivku zvonovitého tvaru.

Má-li studovaná proměnná alespoň přibližně takové normální rozdělení, směrodatná odchylka může začít dělat své divy. Magické operace začínají asi takto. Nejdříve odečteme standardní odchylku od průměru. Pak ji opět přičteme k průměru. Mezi těmito dvěma hodnotami bude vždycky přibližně **68% ze všech pozorování**. Nezáleží na tom, má-li naše křivka tvar profilu hory Řípu, nebo je velice plochá, neboční příkře do výše ve formě jakéhosi falického symbolu, v rozsahu definovaném průměrem ∓ směrodatná odchylka bude vždy 68% pozorování. Když od průměru odečteme a přičteme místo jedné směrodatné odchylky **dvě**, v rozmezí definovaném těmito novými hodnotami bude **95%** pozorování.

A teď si asi říkáte "No, a co z toho?" Kupodivu hodně. Právě tato vlastnost směrodatné odchylky nám umožní dělat některá zajímavá kouzla. Přirozeně, směrodatná odchylka měří homogennost souboru. Umožní nám definovat, jak dobře vypočítaný průměr charakterizuje populaci. Nechá nás formulovat kupř. tvrzení tohoto typu: "Měsíční příjem osob našeho vzorku byl 1.480,- Kčs průměrně. Můžeme předpokládat, že průměrný příjem populace spadá s 95% pravděpodobností do oblasti mezi 1.410,- a 1.550,- Kčs." (Vzpomínáte? O podobných operacích jsme již hovořili v naší kapitole 5. v souvislosti s problémy výběrové chyby.) To je příklad důležité role, kterou směrodatná odchylka hraje pro definování **statistické významnosti** našich výsledků. Díky směrodatné odchylce jsme třeba schopni říci, že existuje jen pětiprocentní pravděpodobnost, že rozdíly v průměrné konzumaci alkoholu ve dvou zkoumaných skupinách jsou náhodné a že pro zbývajících 95% procent můžeme doufat, že rozdíly jsou skutečně funkcí nějaké vlastnosti (třeba pohlaví), podle které byly zkoumané osoby zatříděny do obou skupin.

Tolik tedy o zdánlivě tak jednoduché věci, jako je aritmetický průměr. Jednoduché, ale představující intervalovou statistickou operaci, a to nám umožní aplikovat na ni takové účinné triky jako je měření rozptylu, nebo směrodatnou odchylku. Medián a modus nám to tak lehce neumožní.

V oblasti měření souvislostí jsou rozdíly mezi jednotlivými úrovněmi statistických operací ještě markantnější. Právě z tohoto důvodu je výhodné používat měření na intervalové úrovni co nejčastěji a musíme o tom začít ve výzkumném procesu přemýšlet velmi brzy .

Uveďme si alespoň jeden příklad. Jaká proměnná je "vzdělání"? Nominální, pořadová, nebo intervalová? Jediná odpověď kterou můžeme navrhnout, je: přijde na to, jak jsme vzdělání definovali v naší operační definici. Velice často je vzdělání popsáno v kategoriích jako "neukončené základní vzdělání", "ukončené základní vzdělání" atd. Tedy nejčastěji bude vzdělání patří mezi **pořadové proměnné**. Pro některé účely může být výhodná definice vzdělání jako **nominální proměnné**; kupř. když chceme studovat vliv určitého typu vzdělání na kariéru bývalých studentů: jak se srovnává deset let po dokončení školy plat absolventů průmyslovky s platem absolventů gymnázia, platem absolventů techniky a - nedej Bože - s příjmem studentů s titulem bakaláře v sociologii. Svého času se mnohé doktorské práce studentů v USA zabývaly vlivem prestiže university na budoucí status studentů. Porovnávaly se kupř. platy absolventů nejprestižnějších universit (Big Ten a Ivy League) s platy absolventů jiných univerzit. Mimochodem, často se ukázalo, že zjištěné rozdíly v platech mizí, kontrolujeme-li sociální status studentových rodičů. (Pamatujete si ještě koncept nepravé korelace?)

Hodláme-li použít proměnnou "vzdělání" jako element respondentova sociálního statusu (ale i pro mnohé jiné účely), měli bychom vážně uvažovat o takové operační definici vzdělání, která by nám umožnila jednat se vzděláním jako s intervalovou proměnnou. Je to jednoduché. Proč nepopsat vzdělání jako počet úspěšně dokončených let formálního školního vzdělání? Pro účely mezinárodního srovnávání je to naprosto nezbytné. Čemu v našem vzdělávacím systému odpovídá absolvování "lycea" v Itálii? Čemu odpovídá dokončené středoškolské vzdělání získané v Kanadě? V Ontariu je to 12 nebo 13 let, ve většině jiných provincií 11 nebo 12 let. (A jako by situace nebyla již tak dost komplikovaná, toto středoškolské vzdělání je v Kanadě povinné.) Ale i pro každý jednoduchý docela domácí výzkum je velice výhodné mít vzdělání definováno v letech. Zejména proto, že intervalové proměnné nám, mimo jiné, umožní odpověď na řadu docela zajímavých problémů, jako kupř. "Jak by se typicky zvýšil příjem jedince v závislosti na faktu, že jeho vzdělání vzrostlo o jeden rok, ale všechny ostatní studované faktory, ovlivňující příjem by zůstaly nezměněny?" Nebo: "Co ovlivňuje dosaženou výši jedince více: jeho pohlaví, jeho věk, povolání jeho rodičů?" V naší deváté kapitole uvidíme, že to není tak jednoduché, a že na úrovni intervalového měření mohou mít tyto operace mnohem jednodušší a přehlednější logiku než obdobné operace na nominálních či ordinálních datech.

8.3. Jak moc to souvisí

> Na jednom semináři jsem při odpovědi použila výraz "souvisí jenom malinko" a docent Disman mne požádal, abych nemluvila jazykem hokynářů. Měla jsem říci, "korelace je nízká". Následovala moje obhajoba jazyka hokynářů. Vůbec se mi nelíbilo, že sociologové říkají jednoduché věci co možná nejkomplikovaněji a tak, aby jim skoro nikdo nerozuměl.
>
> *Fotografka Markéta Luskačová v interview publikovaném v My 91, srpen 1991.*

Nepamatuji si tuhle epizodu, ale Markéta má určitě pravdu. Mýlí se jen v jednom bodě. Nebyl to žádný docent, ale docela vykulený a pokorný asistent. Náramně okouzlený divy kvantitativní analýzy v sociálních vědách, tak atraktivní svou zjevnou objektivitou, v době, kdy se značná část sociálních věd omezovala na iterace výroků klasiků tak, aby byly v ještě větším souladu s míněním prezidia ÚV dnes dopoledne. Od té doby ten asistent zestárnul (určitě) a zmoudřel (snad) a dnes tráví mnoho času tím, že přesvědčuje své kanadské studenty, že konečný produkt sociologické práce by měl být napsán tak, aby měl význam i pro ty kouzelné portugalské dědky, co prodávají ryby na Kensington Market v Torontu. Ale to Markéta věděla už dávno, že jedním z důležitých z cílů sociologie je, nebo by mělo být, porozumění. A teď vidím, že tohle by mohl být úvod k naší kapitole o kvalitativním výzkumu. Stejně se tam k Markétě ještě jednou vrátíme.

Už jsme si několikrát řekli, že kvantitativní výzkum je v podstatě jen testování hypotéz, že jeho silnou stránkou je jeho schopnost nám říci, jak moc se mýlíme; k porozumění může kvantitativní výzkum přispět jen někdy, to musí přijít předtím a potom. Testování hypotéz je vlastně produkce výroků o tom, jak silně proměnné souvisí. Tak se teď už podívejme, jak tu souvislost měříme. Že to budeme provádět v číslech, není důležité. Významný je jenom fakt, zda to souvisí hodně, nebo jenom "malinko".

Souvislost jsme si definovali jako přírůstek v pravděpodobnosti uhodnout správně stav jedné proměnné poté co jsme obdrželi informaci o stavu jiné proměnné. Teď jde jen o to, jak nějak chytře vyjádřit tento přírůstek v číslech. Goodman a Kruskal (1954) uvedli velice užitečný koncept, nazvaný **PRE** (Proportional Reduction in Error), což můžeme přeložit jako relativní redukci omylu. Logika tohoto konceptu je půvabně jednoduchá. Můžeme ji vyjádřit třeba takto:

$$PRE = \frac{\text{původní omyl} \quad \text{minus} \quad \text{omyl se znalostí o druhé proměnné}}{\text{původní omyl}}$$

Pro doktora Watsona a ostatní z nás, co nemají rádi vzorečky, můžeme tento koncept vyjádřit prostě v popisu operací, které vedou k výpočtu relativní redukce omylu:

Co uděláme:	Co to znamená:
1. Zjistíme, jak je velký původní omyl v odhadu.	Pamatujete, jak jsme hádali pohlaví osoby, jejíž dotazník byl v obálce? Protože známe distribuci naší proměnné, víme, že bychom asi tak o každém druhém dotazníku hádali mylně.
2. Zjistíme, jak velký bude omyl poté, co jsme obdrželi informaci o druhé proměnné.	To je počet chyb, které uděláme, když víme, zda respondent používá někdy rtěnku.
3. Nový počet omylů odečteme od původního počtu omylů.	To je prostě absolutní velikost úbytku omylu.
4. Získaný rozdíl vydělíme původním počtem omylů.	Tahle operace nám řekne, jaké proporce jsme se zbavili nabytím informace o druhé proměnné.

Ten poslední krok je docela důležitý. Předpokládejme, že znalost nové proměnné vůbec nepřispěje k redukci omylu. To znamená, že počet omylů se znalostí je přesně stejné číslo, jako byl původní omyl. Když tato čísla odečteme, dostaneme nulu a PRE bude tedy nula. Naproti tomu, kdyby odstranila jakoukoliv nejistotu v odhadu hodnoty původní proměnné, počet omylů "se znalostí" by byl nula a její odečtení by množství "původních omylů" vůbe nezměnilo. V závěrečném kroku bychom tedy dělili stejné číslo stejným číslem a tak výsledek by byl jedna. **Ve většině statistických operací měřících sílu souvislostí jednička vyjadřuje perfektní souvislost a nula naprostou nezávislost.**

Naprostá většina měření souvislosti spočívá na logice, podobné logice výpočtu relativní redukce omylu. Liší se však ve volbě strategií, které volí pro optimalizaci odhadu. To je také důvod, proč kroky 2. a 3. jsou dosti nespecifické a nemůžeme je použít přímo pro výpočet.

Podívejme se teď alespoň na jeden příklad toho, jak se to opravdu dělá. Guttmanův koeficient předpověditelnosti, LAMBDA, je nominální měřítko souvislosti. Může tedy být aplikováno na všechny úrovně měření. Důvod, proč jsme ji vybrali, je, že sleduje logiku relativní redukce omylu zcela zřetelně.

Představme si, že máme před sebou horu dvou tisíc a jednoho sta dotazníků v obálkách, obsahujících mimo jiné informaci o respondentově pohlaví a o tom, zda respondent někdy používá rtěnku. Naším úkolem je odhadnout správně pro každý dotazník a udělat to s nejmenším možným počtem omylů. Můžeme třeba dávat na jednu hromadu obálky s dotazníky, o nichž se domníváme, že byly vyplněny ženami, a na druhou ty, o nichž věříme, že je vyplnili muži. Jak minimalizovát počet chyb? LAMBDA používá následující strategii:

Hádej o všech pozorováních, že všechna patří do <u>modální</u> kategorie, t.j. do kategorie s nejvyšší četností pozorování!

Je to trochu divná strategie, ale za chvíli jí přijdete na chuť. Informaci o distribuci našich dvou proměnných nalezneme v následující tabulce 8.5.

Tabulka 8.5.

Pohlaví:

Používání rtěnky:

	MUŽ	ŽENA	Celkem:
ANO	50	900	950
NE	950	200	1150
Celkem:	1000	1100	2100

Na začátku nevíme nic jiného, než informaci o rozložení vzorku podle pohlaví, tedy kolik je mužů a kolik žen v našem vzorku. Formálně řečeno, známe marginální (okrajové) četnosti pro proměnnou, kterou se snažíme uhodnout. V našem případě jsou to sloupcové marginální četnosti, součty sloupců vytištěné v posledním řádku tabulky. Víme tedy, že ve vzorku je 1000 mužů a 1100 žen.

A nyní už můžeme postupovat podle strategie, kterou LAMBDA doporučuje pro optimalizaci odhadu. V našem vzorku je více žen než mužů; tedy kategorie "žena" je modální kategorií pro proměnnou "pohlaví". Dáme tedy všechny obálky na jednu hromadu a prohlásíme, že všechny dotazníky byly zodpovězeny ženami.

Dr. Watson:

Ale to je nesmysl. Tisíc dotazníků bude zařazeno zaručeně špatně!

Dr. Watson má pravdu. Jenže zároveň máme také jistotu, že pro každý dotazník vyplněný ženou, a těch je více, jsme odpověděli správně. Ale v každém případě vidíme, že **původní počet omylů = 1.000.**

A kolik omylů uděláme v odhadu pohlaví, když máme k dispozici informaci o druhé proměnné, o používání rtěnky? Nyní začneme odhadovat pohlaví zvlášť pro ty, kteří používají někdy rtěnky, a zvlášť pro ty, kteří ji nepoužívají. Východiskem jsou teď dvě hromady dotazníků. Začneme třídit jednu, třeba ty, co používají rtěnku. Potřebnou informaci najdeme v prvém řádku tabulky. Je zde 5O mužů a 900 žen. Zařadíme všechny dotazníky do modální kategorie. Dostaneme tak **50 nových omylů**, protože 50 mužů bylo chybně zařazeno mezi ženy.

Ale to nejsou ještě všechny nové omyly. Musíme ještě roztřídit skupinu respondentů, kteří nepoužívají rtěnku. Data v druhém řádku tabulky jasně ukazují, že pro tuto skupinu modální kategorie jsou "muži", kterých je 950. Zařadíme tedy všech 1150 dotazníků do této kategorie a tím vyprodukujeme **dalších 200 nových omylů.**

A teď už známe všechna čísla, která potřebujeme pro výpočet našeho koeficientu:

Počet původních omylů: počet mužů mylně zařazených mezi ženy na základě sloupcových marginálních četností	1000
Počet nových omylů: počet mužů mylně zařazených mezi ženy ve skupině používající rtěnku (50) + počet žen mylně zařazených mezi muže ve skupině nepoužívající rtěnku (200)	250

Absolutní úbytek v počtu omylů: původní omyly mínus všechny nové omyly	750
Relativní úbytek v počtu omylů: absolutní úbytek v počtu omylů vydělený původním počtem omylů	LAMBDA= 0.75

Tak a teď už umíme vypočítat alespoň jeden z koeficientů souvislosti. Avšak mnohem důležitější je, že rozumíme logice tohoto výpočtu. Nadto v našem případě koeficientu LAMBDA můžeme zcela intuivně rozumět tomu, co znamená jeho velikost: vypočítaná hodnota .75 nám říká, že znalost o tom, zda respondent užívá rtěnku, zmenšila o tři čtvrtiny počet omylů v odhadu pohlaví respondentů. Bohužel, ne u všech operací měřících souvislost, je taková jasná a intuitivní interpretace koeficientů možná. A mimochodem, neočekávejte ve studiu sociálních vztahů příliš vysoké koeficienty. Pro mě je každý koeficient vyšší než .30 dobrým důvodem k oslavě.

Cvičení 8.4.

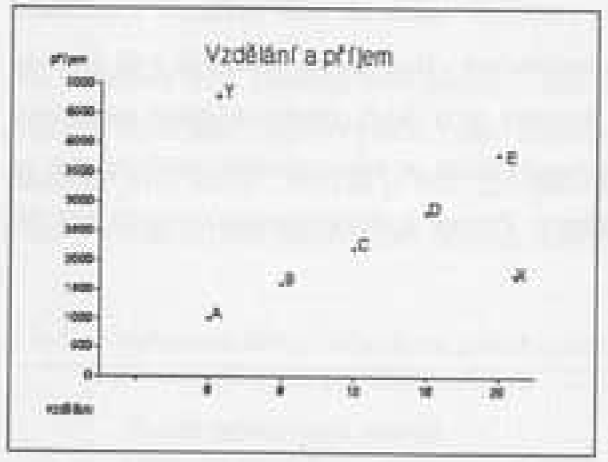
Zkuste teď obrátit naši úlohu. Vypočítejte lambdu pro takovou situaci, kdy odhadujeme, kteří respondenti používají rtěnku a informace o jejich pohlaví nám má pomoci v tomto úkolu. Všechna data jsou opět v tabulce 8.5.

Jakmile jste si uvědomili, že musíme teď používat sloupce místo řádků, tak to nebylo těžké, že? Zajímavé je, že nový koeficient .737 je o něco nižší, než byl náš původní. Lambda je totiž **asymetrický koeficient**. Pro mnoho jiných měření souvislosti, jako kupř. pro korelační koeficient, je jedno, zda X je nezávislá a Y závislá proměnná, nebo zda je tomu naopak. U takových symetrických statistik dostaneme v obou případech stejný koeficient.

8.4. Souvislosti mezi aristokraty

Intervalové proměnné jsou opravdu aristokraty mezi ostatními proměnnými. Můžeme s nimi provádět mnoho operací, které jsou pro jiné úrovně měření problematické, nebo nemožné. Proč je tomu tak? Podívejme se na to podrobněji.

Máme-li dvě intervalové proměnné, můžeme každého jedince znázornit jako bod v dvojrozměrném prostoru. Poloha tohoto bodu pak charakterizuje hodnotu, kterou mají obě z proměnných pro dotyčného jedince. Tohle zní mnohem složitěji, než to ve skutečnosti je. Podívejme se na jednoduchý příklad a na obrázek.

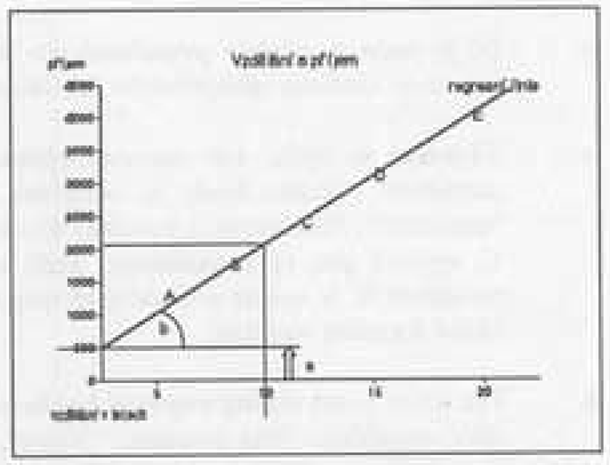


Graf 8.1.

V grafu 8.1. vodorovná osa X představuje vzdělání popsané v letech, kolmá osa Y reprezentuje hrubý měsíční příjem v korunách. O každém jedinci reprezentovaném na našem grafu můžeme říci, jaké je jeho vzdělání a jaký je jeho příjem. Tak jedinec A má 6 let vzdělání a měsíční příjem tisíc korun. Pro respondenty B až E platí roste rovnoměrně se vzděláním; tak kupř. respondent E má dvacet let vzdělání a příjem 4.000 Kčs. Pak tu máme ještě dvě anomálie: Respondent Y má nízké vzdělání a velice vysoký příjem. Mohl by to být třeba vekslák. A co můžete říci o jedinci X? Prosím, nekřičte moc hlasitě, že je to sociolog nebo něco podobného.

A teď si představme, že bychom se zbavili obou těch úchylkářů a v našem vzorku by zůstali jen ti ideální respondenti, u nichž s přirůstajícím vzděláním rovnoměrně přirůstá i příjem. Pak bychom těmi ukázněnými pozorováními mohli protáhnout krásnou přímku. To je ta silná, šikmá čára na našem grafu 8.2. Říká se jí **regresní přímka**. Teď bychom mohli

Graf 8.2.



začít provádět úplná kouzla. Regresní analýza nám umožní pro každého jedince určit hodnotu proměnné Y, když pro něj známe hodnotu proměnné X. V situaci znázorněné v našem grafu bychom to mohli udělat naprosto přesně a bez jakéhokoliv počítání: od známé hodnoty vzdělání povedeme kolmou čáru až k bodu, kde se dotkne regresní linie. Pak zahneme v pravém úhlu doleva a naše přímka protne osu Y v místě odpovídajícím jednotlivcově příjmu.

Dr.Watson:

Ale proč bychom to dělali? Vždyť pro každého našeho jedince známe jeho vzdělání i jeho příjem. Jinak bychom nebyli schopni nakreslit ho na správné místo na naší mapě.

V téhle kapitole má dr. Watson často pravdu. Alespoň částečně. Ono nám o samotné hádání zatím nejde, i když v příští kapitole uvidíme, že to může někdy být velice užitečné. Zatím můžeme alespoň naznačit, že pravidla pro předpověď hodnoty Y na základě znalosti hodnoty X je možno generalizovat na populaci, kterou vzorek reprezentuje. V našem vzorku není nikdo, kdo by měl právě 10 let vzdělání. Nicméně jsme schopni předpovědět pro kohokoliv z populace, jaký by byl jeho příjem. Tato operace je znázorněna kolmou čárou, která v grafu 8.2. protíná desítku na ose X.

Tuhle operaci můžeme také udělat bez jakéhokoliv kreslení, prostě matematicky. Tady je pro to vzoreček. Uvádíme jej hlavně proto, abychom ukázali, že algebra nekouše.

$$y = a + bx$$

Co je co:

y — To je hodnota **závisle proměnné** pro daného jednotlivce. Závisle proměnná je ta, kterou se snažíme předpovědět na základě naší znalosti o nezávisle proměnné X.

a — Všimněte si bodu, kde regresní přímka protíná kolmou souřadnici Y. Hodnota proměnné v tomto bodu je nazývána **konstanta** (v anglicky psané literatuře "intercept"). Můžeme si ji vysvětlit docela jednoduše: je to hodnota závislé proměnné Y, typická pro ty respondenty, kteří mají nejnižší možnou hodnotu v nezávisle proměnné X. V našem případě je to tedy typický příjem těch jednotlivců, kteří nemají žádné formální vzdělání.

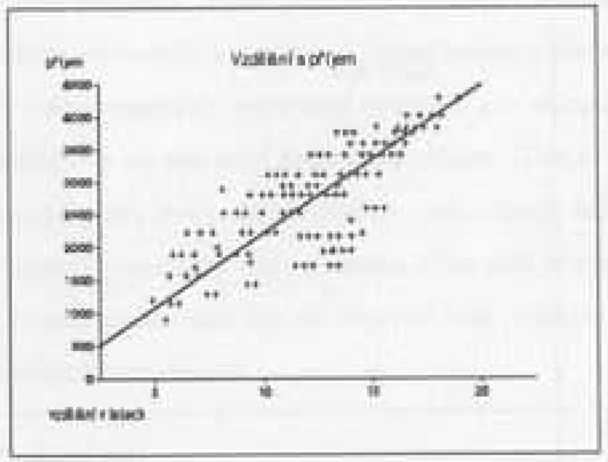
b — Tak tohle je ten slavný **regresní koeficient**. Angličtina pro něj používá jméno, které dost vysvětluje jeho podstatu: "slope", to je "svah". Čím více přibývá příjem s rostoucím vzděláním, tím příkřejší je sklon regresní linie. Regresní koeficient je v podstatě informace, o kolik vzroste Y když nezávisle proměnná vyrostla o jednu jednotku. Jinak řečeno, regresní koeficient nám řekne, o kolik korun vzroste příjem, když vzdělání vzrostlo o jeden rok.(Regresní koeficient je v grafu znázorněn úhlem b.)

x — Tohle je opravdu jednoduché. Je to hodnota **nezávisle proměnné**, pozorovaná pro daného jedince; tedy v našem případě délka jeho vzdělání.

A tady ještě přidáme instrukci pro dr. Watsona, jak odhadnout příjem, když známe vzdělání:

1. Každému v našem vzorku dáme stejnou konstantní sumu (konstantu, "intercept") odpovídající příjmu osoby s nejmenším možným vzděláním.
2. Kromě toho každý jednotlivec dostane prémii, závisející na tom, kolik má let vzdělání. (Tato individuální prémie je částka, odpovídající počtu let vzdělání, vynásobenému regresním koeficientem.)

To není těžké, že? Jenže ve skutečnosti je to trochu jinak. Takhle přesně bychom mohli odhadovat váhu kolejnice z její délky, ale ne příjem na základě vzdělání. Ve studiu sociálních jevů nikdy nedostaneme taková data, která by se nalézala přesně na regresní přímce, ale data hodně rozptýlená. Když máme hodně štěstí, data mohou být distribuována tak, jako v grafu 8.3.

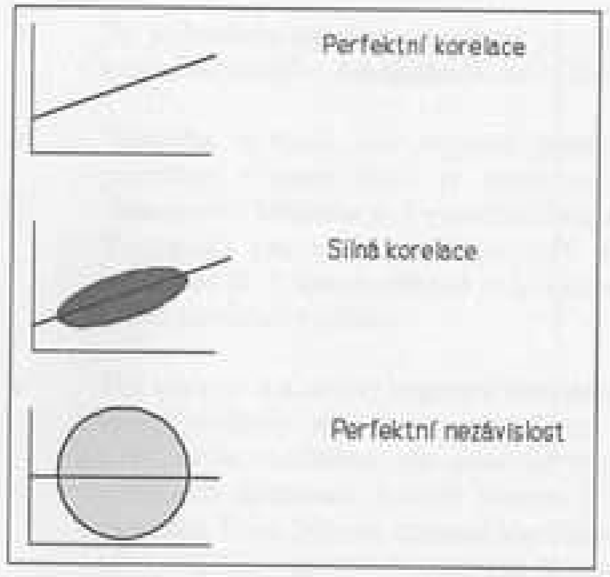


Graf 8.3.

Dovolte, abych citoval Dismanův první zákon: "Data jsou potvory." Ne, že by chtěla, spíš musí. Sociálním datům nezbývá nic jiného. Kdybychom se zabývali souvislostí mezi délkou a váhou kolejnic, nic - snad jen s vyjímkou hrubé výrobní chyby - by nemohlo podstatně ohrozit správnost našeho odhadu. Délka je jediný faktor, který ovlivňuje váhu.

S příjmem, jako s každým sociálním jevem, je to jinak. Příjem je ovlivňován desítkami různých faktorů a my jsme zde **zredukovali** všechny tyto faktory na jediný, na vliv vzdělání, a tak někdo s vysokým vzděláním má jen krátkou pracovní zkušenost. Proto je jeho příjem nižší, než odpovídá jeho vzdělání, a na grafu se objeví pod regresní linií. Někdo s poměrně nízkým vzděláním může mít dlouhou praktickou zkušenost, pracuje nadto v preferovaném povolání, a tak se objeví vysoko nad regresní přímkou. Je to opět to naše staré strašidlo, které nás provází na každém kroku výzkumu: nevyhnutelná redukce informace v nezvládnutelně velkých "přirozených systémech".

Nicméně všechny postupy, které jsme navrhli pro odhad příjmu na základě naší znalosti o vzdělání, stále platí. Rozdíl je jen v tom, že náš odhad příjmu bude daleko méně přesný. v grafu 8.3. vidíte, že i zde je regresní linie. Tentokrát však není tato přímka proložena všemi pozorováními, ale **je lokalizována tak, aby průměrná vzdálenost od této přímky byla co nejmenší.**



Graf 8.4.

Postup pro výpočet odhadnutého příjmu je stejný, jaký jsme navrhli dříve. Předpověď bude však přesná jen pro ty jedince, kteří se nalézají přesně na regresní linii. **Čím je vzdálenost jedince od regresní linie větší, tím vyšší bude náš omyl v odhadu.** A tady jsme u logického základu, na němž je vybudováno měření souvislosti mezi dvěma intervalovými proměnnými. Můžeme si to opět nejlépe vyjádřit graficky: Všechna pozorování na naší mapě mohou být uzavřena do křivky. V anglicky psané literatuře se této křivce říká "envelope" (obálka). Když jsou pozorování rozptýlena po ploše našeho grafu náhodně, obálka bude mít formu kruhu.

V takové situaci nemáme žádné vodítko, jakým směrem regresní linii táhnout. Znalost o proměnné X nezlepší naši schopnost odhadnout hodnotu proměnné Y. Jsou-li pozorování nakupena v nějakém užším prostoru, obálka bude mít tvar elipsy. Regresní linie bude shodná s dlouhou osou elipsy, a čím je elipsa užší, tím méně se budeme mýlit v našem odhadu. Až se obě strany elipsy překryjí, jako v gumičce, kterou jsme natáhli, z elipsy se stane čára a náš odhad bude zcela bez omylů. To všechno jasně vidíme v našem grafu 8.4.

Tahle metafora dobře odpovídá konceptu **korelačního koeficientu.** Jeho logiku si opět můžeme vysvětlit v termínech, které známe, v termínech redukce omylu. Neznáme-li hodnotu X (tedy vzdělání), optimální strategie pro minimalizaci omylu bude předstírat, že všichni jedinci ve vzorku mají **průměrný** příjem. (Ten je reprezentován vodorovnou linií v poslední kresbě grafu 8.4.) Má-li obálka tvar elipsy, můžeme pro odhad závisle proměnné použít regresní rovnici, kterou už známe. Čím užší je elipsa, tím přesnější bude náš odhad. Jsou-li všechna pozorování jen na regresní linii, velikost omylu klesla na nulu a máme zde případ perfektní souvislosti.

Cvičení 8.5.
Představte si, že regresní linie v první kresbě v grafu 8.4., to je, začíná vysoko u osy Y a klesá doprava, dolů k ose X. Jakou relaci takové regresní linie reprezentuje?

Korelační koeficient (zpravidla symbolizovaný písmenem r) může nabývat hodnoty mezi 1 a -1:

r = 1	**perfektní pozitivní korelace** S rostoucí hodnotou X hodnota Y vzrůstá. Hodnotu Y odhadneme na základě znalosti hodnoty X bez jakéhokoliv omylu.
r = 0	**naprostá nezávislost** Znalost hodnoty X nezlepší naši schopnost odhadnout správně hodnotu Y.
r = -1	**perfektní negativní korelace** S přirůstající hodnotou X hodnota Y klesá. Hodnotu Y odhadneme na základě znalosti hodnoty X bez jakéhokoliv omylu.

Korelační koeficient r má jednu velice příjemnou vlastnost. Jeho algebraická definice je lehce pochopitelná intuitivně

r^2 = proporce variance v proměnné Y, vysvětlitelná změnami v X

Totéž můžeme vyjádřit sice méně přesně, ale, doufejme, jasněji, asi takto: Korelační koeficient vynásobený sám sebou nám poskytne informaci, jaké procento rozdílů existujících v příjmu se zdá být vysvětlitelné rozdíly, které existují ve vzdělání. To slovo "zdá" tu zdůrazňujeme zcela záměrně. K tomu se později ještě vrátíme.

A teď si v kostce zopakujme to nejdůležitější:

Regresní koeficient (b)

nám řekne, co máme hádat.

Korelační koeficient (r)

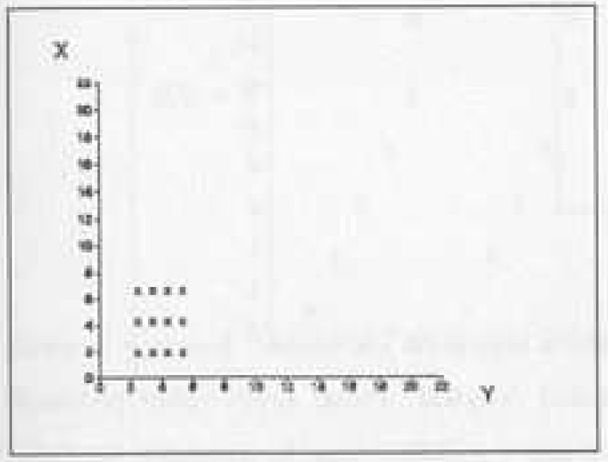
jak dobře budeme schopni hádat.

Nejzajímavější a pro nás nejdůležitější je aplikace korelační a regresní analýzy v takových situacích, kdy studujeme několik proměnných současně. Pro společenské vědy, o nichž jsme si jen s lehkým přeháněním řekli, že v nich souvisí všechno téměř se vším ostatním, je tato schopnost regresní analýzy tak důležitá, že jí věnujeme celou kapitolu.

Jistě jste si už všimli, že už známe regresní rovnici, návod jak vypočítat Y, známe-li X, ale nevíme zatím vůbec, jak vypočítat korelační a regresní koeficient. Já vám návod k provedení těchto operací prostě zatajím. Ty výpočty nejsou těžké, ale jsou hodně pracné, časově náročné. Naše knížka nám neposkytuje dost místa pro jednoduché vysvětlení. Dnes by snad jen vězeň v izolaci vypočítával korelační a regresní koeficienty ručně. I nejhloupější počítač a každá trochu lepší kalkulačka to dovedou udělat velice rychle.

Jenže v tom je právě háček! Ono to s tou korelací a regresí není vždycky tak jednoduché. Nemůže za to ani tak ta analýza, ale data mohou zase jednou být potvory; mohou mít někdy takovou konfiguraci, která nás neomylně zavede k falešným závěrům. Pokud byste se chtěli

sami pustit do takové analýzy, bylo by dobré poradit se s někým, kdo umí **opravdovou** statistiku. Tady vám předvedu jen některé z léček, které na nás konfigurace dat může nastrojit.



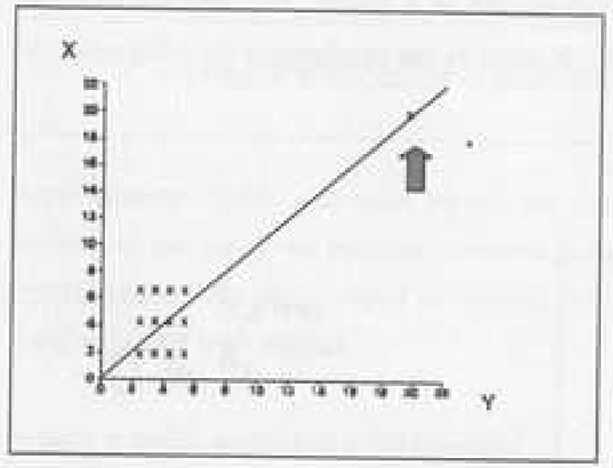
Graf 8.5.

$r^2 = 0$

$b = 0$

V grafu 8.5. vidíme jednoduché seskupení 16 pozorování. Zamyslete se nad nimi: existuje nějaká souvislost mezi proměnnými X a Y? Odpověď je docela jasná: bez ohledu na to, jaká byla pozorované hodnota X, jedinec bude mít v proměnné Y hodnotu 2, nebo 4, nebo 6. Je zřejmé, že graf symbolizuje situaci perfektní nezávislosti. Optimální strategie pro odhad hodnoty Y je navrhnout, že všechna pozorování mají hodnotu Y rovnou průměru této proměnné.

A teď se podívejme, co se stane, když k našim šestnácti pozorováním přidáme jedno další, s vysokými hodnotami obou proměnných. Toho přidaného jedince jsme označili v grafu 8.6 šipkou.



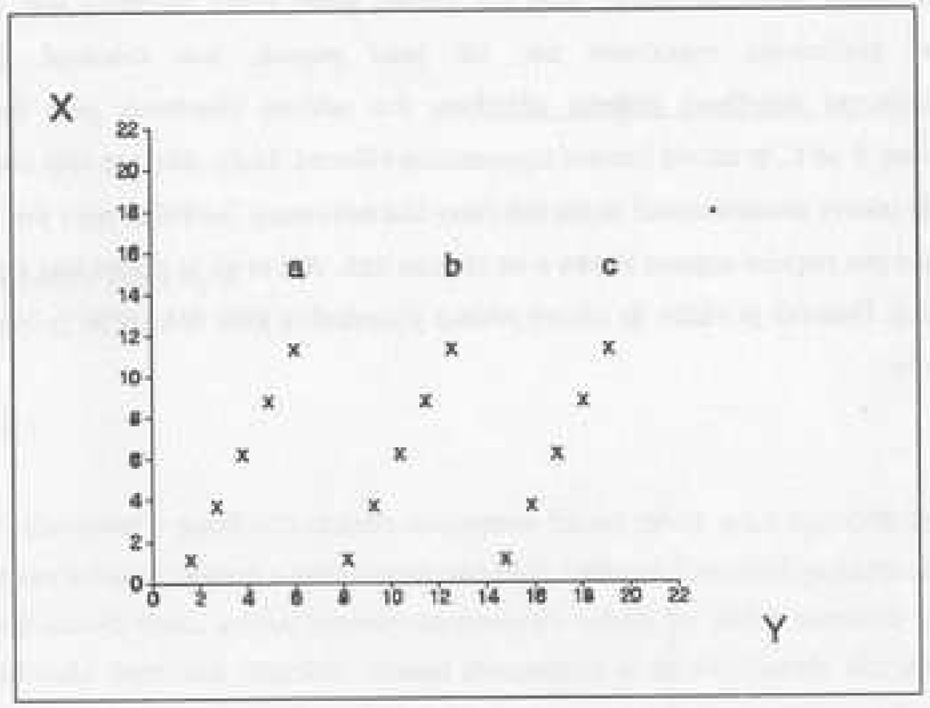
Graf 8.6.

$r^2 = .878$

$b = .931$

Podívejte se teď na nové hodnoty korelačního a regresního koeficientu. Korelace je téměř perfektní. Co vlastně způsobil ten jeden, jediný úchylkář? Prostě zvýšil velice podstatně rozptyl našeho vzorku. Matematicky je tu všechno v pořádku; víme, že umocněný korelační koeficient odpovídá proporci rozptylu v závisle proměnné, kterou je možno vysvětlit rozdíly v proměnné X. Ne tak docela v pořádku je vše na úrovni interpretace dat. Téměř všechen rozptyl byl vnesen do našeho vzorku tím jediným, novým pozorováním. Ta velká vysvětlující síla r^2 se týká jenom tohoto úchylkáře ve vztahu ke zbytku pozorování. Oba nové koeficienty nám prakticky vůbec nepomohou k lepšímu porozumění vztahů v jádru našeho vzorku, v původních našich 16 pozorováních.

Ale teď se pro změnu podívejme na spíše optimistický příklad. Distribuce v grafu 8.7. naznačuje, že obě proměnné jsou prakticky nezávislé. Ovšem, všimli jste si už, že data mají zajímavou konfiguraci, kterou můžeme dobře využít. Rozdělíme prostě naši původní populaci do tří. V populaci A budou všechna pozorování, která mají hodnotu X menší než 8. V populaci budou jednotlivci s X mezi 8 a 12 a v populaci C budou všechna ostatní pozorování. A teď vypočítáme pro každou populaci zvlášť korelační koeficient. Vlastně ani nemusíme počítat. Na prvý pohled vidíme, že v každé populaci jsou všechna pozorování přesně na regresní linii, a tedy v každé ze tří subpopulací existuje perfektní souvislost mezi X a Y.



Graf 8.7.

Cvičení 8.6.

Tak tohle je trochu těžší cvičení. Použijte svoji představivost a navrhněte takové rozložení dat, které by v celé populaci představovalo silnou <u>pozitivní</u> souvislost a v subpopulacích silnou <u>negativní</u> souvislost.

Co se tu vlastně stalo? Ilustrovali jsme zde vlastně jednu velice důležitou věc: korelační a regresní koeficienty vypočítané tak, jak jsme popsali, jsou **lineární**. Snaží se **charakterizovat distribuci jedinou přímkou.** Pro některé distribuce, jako kupř. naše subpopulace A až C, je taková lineární reprezentace výborná. Jindy, jako pro celý obsah grafu 8.7., může takový lineární model ztratit důležitou část informace. Statistika zná i jiné postupy, používající pro popsání regrese křivku a ne přímou linii. Ale to už je trochu nad možnostmi naší knížky. Důležité je vědět, že takový přístup je možný, a ještě důležitější je být zadobře se statistiky.

Řešení úkolů z kapitoly 8.

Cvičení 8.1.

Perfektní souvislost v sociálních datech dostaneme jenom tehdy, když vypočítáme korelaci mezi respondentovým věkem a jeho rokem narození, nebo když omylem požádáme počítač, aby vytiskl tabulku třídící určitou proměnnou samu se sebou. Ve všech ostatních případech se projeví naše chronická choroba: velikost přirozených systémů. Desítky různých faktorů mohou ovlivňovat respondentovy postoje, a my jsme zredukovali celou složitou síť vztahů na měření třeba toho, jak je silný vliv vzdělání. Náš výpočet není kontrolován pro takové faktory jako věk, zdravotní stav atd., a i kdyby vliv vzdělání měl určující charakter, vypočítaný koeficient se vlivem těchto dalších faktorů může projevit jako docela slabý.

Cvičení 8.2.

To by vyjadřovalo opět perfektní, ale **negativní** souvislost.

Cvičení 8.3.

Pro populaci - a je to velice snadné - i z hlavy snadno určíme, že rozptyl i směrodatná odchylka musí být 0. Pro populaci B už musíme trošku počítat, abychom zjistili, že rozptyl je 1280 a směrodatná odchylka je 16.

Cvičení 8.4.

Postup výpočtu je téměř shodný jako předtím. Jediný rozdíl je v tom, že tentokrát nepoužijeme řádky, ale sloupce tabulky. Z posledního sloupce - z marginálních četností - vypočteme počet starých omylů, ze sloupců v poli tabulky vypočteme počet nových. Zbytek operací už známe. Konečný výsledek bude poněkud nižší než náš původní:

LAMBDA = .737

Lambda je asymetrický koeficient. Jeho hodnota, předpovídáme-li X na základě Y, se může někdy docela dramaticky lišit od hodnoty vypočítané pro předpověď Y na základě naší znalosti o X.

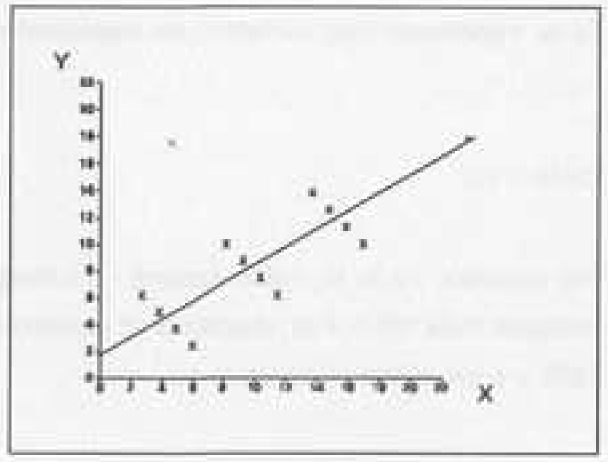
Cvičení 8.5.

To už známe: to je perfektní negativní vztah. "Čím vyšší vzdělání, tím nižší příjem." Pro mne zní tenhle výrok nepříjemně realisticky.

Cvičení 8.6.

Podmínkám našeho úkolu by mohl odpovídat třeba tento graf:

Graf 8.8.



Konfigurace dat ukazuje, že v celém souboru existuje **silná pozitivní relace** mezi hodnotami X a Y. Naproti tomu v každém podsouboru můžeme pozorovat **perfektní negativní souvislost**.